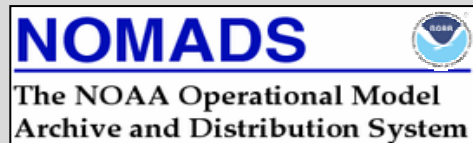


**The  
NOAA  
Operational Model Archive and Distribution System (NOMADS)  
Program Plan**

**and  
Data Management Vision**



**A Data Management Plan for use by NOMADS Participants and  
by NOAA IT Managers**

**DRAFT**

**NOAA Core Collaborators**

Climate Diagnostic Center  
Forecast System Laboratory  
Geophysical Fluid Dynamics Laboratory  
National Climatic Data Center  
National Centers for Environmental Prediction  
Pacific Marine Environmental Laboratory

**Core External Collaborators**

COLA  
DOE/LLNL/PCMDI  
NASA/GCMD  
UCAR/NCAR  
UCAR/Unidata  
BADC  
BOM

**Glenn K. Rutledge, Principle Investigator**

**January 12, 2003**

## Executive Summary

Climate study is fundamentally multidisciplinary. As we strive to understand its complexity, researchers from different fields and different locations must become engaged in large multinational teams to tackle grand challenge problems. What is needed is a software infrastructure to support this multidisciplinary virtual organization, across Agencies and institutions. This document is intended for use by both NOAA management, to advance NOMADS and NOMADS-related activities, and for scientists- the users of these data and framework for the advancement of our understanding of our earth system.

Over the past two years, the National Climatic Data Center (NCDC) has developed new partnerships extending across organizational boundaries. The effort described within this program plan was years in the making. It is just now that visionaries across the geo-sciences are forming partnerships that will provide community code and tools that support collaboration and data sharing with location-independent equal-access to shared resources (data, visualization, supercomputers, experiments, whiteboard, etc.). This document serves as an introduction to, and as a template, for NOMADS participants to understand and advance a new distributed data service within their respective Agencies. The NOAA Operational Model Archive and Distribution System (NOMADS) is just such a distributed service that is being developed as a unified climate and weather archive so that users can make decisions about their specific needs on time scales from days (weather), to months (El Nino), to decades (global change).

As supercomputers increase the temporal and spatial resolution of models, and demands for on-line access to large-array data increase, current communication technologies and data management techniques are inadequate. Rapid development in communications technologies over the last decade has been essential to provide significant gains in our ability to access our national data archives. These technologies have been expected to enable collaborative research between institutions to rapidly grow. However, the difficulties in achieving accessibility and interoperability in these national archives have delayed the realization of these potential gains. It is clear that it is no longer sufficient for any one national center or laboratory to develop its data services alone. Both researchers and policy-makers alike now expect national data assets to be easily accessible and interoperable with each other, regardless of their physical location. As a result, an effective interagency distributed data service requires coordination of data infrastructure and management extending beyond traditional organizational boundaries.

The primary national responsibility for the archive and long-term stewardship of climate and weather data rests with the NCDC. However, there exists today no long-term archive for climate and weather models. To address the need for distributed access to Numerical Weather Prediction (NWP) and General Circulation Models (GCM) and data, the National Climatic Data Center (NCDC), along with the National Centers for Environmental Prediction (NCEP) and the Geophysical Fluid Dynamics Laboratory (GFDL), initiated the infrastructure project NOMADS. NOMADS addresses data access needs as outlined in the U.S. Weather Research Program (USWRP) Implementation Plan for Research in Quantitative Precipitation Forecasting and Data Assimilation to "redeem practical value of research findings and facilitate their transfer into operations." NOMADS is a major collaborative pilot effort spanning multiple Government

agencies (US, Australia, UK, and Europe) and academic institutions. NOMADS overcomes the barriers for interoperability by using established and emerging technologies and data transport conventions, but most importantly- partnerships, to access and integrate model and other data stored in geographically distributed repositories in heterogeneous formats. A more detailed NOMADS Program Plan needs to be generated as an update to this document. This will be developed in conjunction with NOMADS co-PI's and others involved in the NOMADS effort during the review of this document.

NOAA has an opportunity to be a the forefront of this new data access philosophy, and must now provide Agency level support to fuse this technology, and science-based approach into mainstream operations.

## PREFACE

Under President Bush's Climate Change Research Initiative (CCRI), access to and understanding of climate models is a high priority. The primary national responsibility for the archive and long-term stewardship of climate and weather models rests with the NCDC. However, there exists today no long-term archive of models and associated data. Because of this, national and international capabilities to develop systematic approaches to climate model evaluation, and the reduction of uncertainties, are severely hampered. NOMADS is a working example that fulfills the goals as outlined by the CCRI, Chapter 12 "Grand Challenges in Modeling, Observations, and Information Systems" by providing format independent access to heterogeneous data across multiple disciplines and institutions. The current success of NOMADS can be attributed in part, toward the advancement of virtually all of the elements within two major themes in the NOAA Strategic Plan: 1) Environmental Assessment and Prediction Mission; and 2) National Capabilities and Supporting Infrastructure. The NOMADS framework facilitates climate model and observational data inter-comparison issues as discussed in documents such as the Intergovernmental Panel on Climate Change (IPCC 1990, 1995, 2001) and the U.S. National Assessment (2000).

It is well known that a major obstacle for climate change detection and attribution studies is the lack of systematic long-term approaches to model evaluation. The NOMADS software infrastructure is a framework that promotes coordinated approaches to model evaluation. With NOMADS and easy access to multi-institutional, multi-discipline coupled model results (land, atmospheric, ocean & sea ice, carbon cycle, etc.), carefully designed systematic model intercomparisons are greatly facilitated. NOMADS does not and will not address approaches to climate and weather model evaluation, however it does lay the framework for this to occur.

The fundamental issue that this program seeks to address is how NOAA and its partners can organize its distributed climate and weather models and data into a cohesive presence and perform real-time and retrospective climate and weather model analysis and inter-comparisons. This will for the first time allow NOAA scientists access to archived models and data to verify and improve climate change and detection processes, improve short-term and seasonal forecasts, and long-term global climate simulations under a distributed client-server framework. NOMADS will also allow users at any level, to obtain weather and climate information for users to make better, informed decisions about how nature will impact their future, either in their life or in their business decisions.

The Kyoto protocol has clearly demonstrated the will of nations to take remedial action to mitigate the impact of anthropogenic climate change. This places a heavy responsibility on the climate research community to provide more reliable computations of the anticipated climate change for alternative scenarios of greenhouse gas emissions. Significantly improved assessments of the details of climate change and its impacts will be required, particularly at the regional level. This in turn will require an improved understanding of the climate system and its interactions with the socio-economic system. As we are unable to experiment with a single system Earth, modeling is the only analytical tool available for understanding the dynamics of the climate system and predicting the future evolution of climate, either under natural conditions

or under the influence of anthropogenic impacts. However, the true coupling of models (atmosphere-ocean; ocean-cyrospheric; carbon cycle; etc.) and intercomparisons of model results, within and between research facilities on an international scale does not currently exist. The NOMADS team are defining required tools, and in terms of science-based requirements and planning, and a series of workshops and informal meetings are on going.

Because of its bottom up approach and distributed data access philosophy, NOMADS now enjoys a strong DOE, NASA, NOAA, NSF, and University partnership. This program plan is the result of many man-years of development, both in terms of software development, and visionaries across multiple organizations. NOAA has the opportunity to be at the forefront of these new developments in data access and earth system understandings. Without NOAA level support however, this opportunity may be lost. Processes to implement NOMADS architectures within NOAA operational environments are now required. This document will help educate Information Technology (IT) managers and NOAA management to achieve this goal.

# NOMADS Program Plan

## Contents

Section	Description	Page
	Executive Summary	
	Forward	
	Acronyms	
	Tables and Figures	
1.0	Background	
2.0	A New Framework	
3.0	Purpose and Scope	
4.0	Program Goals and Objective- An Overview	
4.1	NOMADS Program Goals	
5.0	NOMADS Operating Principles	
6.0	Roles and Responsibilities	
6.1	NCDC Participation	
6.2	NCDC Deliverables- FY03	
6.3	NCDC NOMADS to HDSS Mass Storage Interface	
7.0	The NOMADS Framework- Participation Overview	
8.0	Teams and Working Groups	
8.1	Steering Group	
8.2	Science Team	
8.2.1	Science Based Planning	
8.3	Technical Team	
8.4	Web Portal Development Team	
9.0	NOMADS Initial System Operating Capability – V0.1	
10.0	Templates for Participation	
10.1	Data Provider	
10.1.1	Conventional Data Access Methods	
10.2	Data User	
10.2.1	OpENDAP	
10.2.2	NOMADS Desktop Services- Clients	
10.2.2.1	Client Libraries	
10.3	NOMADS Servers	
10.3.1	GrADS-Data Server	
10.3.2	Climate Data Analysis Tools	
11.0	Data Cataloging	
11.1	THREDDS	
11.2	GMU	
11.3	NASA/GCMD	
12.0	Grid Services and NOMADS	
13.0	Funding Approaches	

13.1	Operational Staffing Requirements
14.0	External Collaborations
15.0	NOMADS Participants and Data Suppliers
15.1	Potential Data Suppliers
15.2	International
16.0	Application Programming Interfaces
17.0	Communications Protocols
18.0	Data Conventions
19.0	Collaborators
20.0	Data Availability-
20.1	NCEP
20.1.1	Real-Time Services at NCEP
20.1.2	NCEP Model Input Data
20.1.3	Introduction to 4DDA
20.1.4	Model ReRun and Retrospective Capability
20.1.5	Direct ftp Services
20.1.6	NCEP Model Volume Requirements
20.1.7	WRF Model
20.2	NCDC
20.2.1	NOAAPort
20.2.2	NCEP Regional Reanalysis
20.2.3	Full Suite NCEP Global and Meso
20.2.4	GDAS
20.3	GFDL
20.4	NCAR
20.5	CDC
20.6	COLA
20.7	NASA
20.8	LLNL/PCMDI
21.0	The Future of NOMADS
21.1	Future Requirements
21.2	US National “HelpDesk”
22.0	Bibliography

## **Part Two**

### **A Data Management Vision under the NOMADS Philosophy**

1. A Vision for Data Management and Usability
2. Achieving the Vision with Existing Resources
3. Strategy
  - a. NOMADS Development Strategy
  - b. Metadata Conventions
4. Areas of Development
  - a. Data Cataloging
5. Clients for Scientists
6. Data Discovery

- 7. Data Access**
- 8. Data Archiving and Preservation**
- 9. Data Coordination and Training**
  - a. Data Providers**
  - b. Data Service Developers**
  - c. Display and Analysis Tool Developers**
- 10. OPeNDAP Overview**
  - a. Metadata**
  - b. The Web and OPeNDAP**

## Acronyms

ACARS	Aircraft Communications and Reporting System
AOGCM	Atmospheric-Ocean General Circulation Model
API	Application Program Interface
AMIP	Atmospheric Model Intercomparison Project
AS	DODS/OPENDAP Aggregation Server
ASCII	American Standard Code for Information Interchange
AVN	NCEP Aviation Model
AWIPS	NWS Advanced Weather Interactive Processing System
BADC	British Atmospheric Data Center
BUFR	Binary Universal Form for the Representation of meteorological data
CAS	NCEP Cycling Analysis System
CCRI	Climate Change Research Initiative
CCSM	NCAR Community Climate System Model
CDAT	PCMDI Climate Data Analysis Tools
CDC	NOAA Climate Diagnostics Center
CDP	NCAR Community Data Portal
CGI	Common Gateway Interface
CLASS	NESDIS Comprehensive Large Array Stewardship System
CMIP	Coupled Model Intercomparison Project
COARDS	Cooperative Ocean-Atmosphere Research Data Standard
COLA	Center for Ocean-Land-Atmosphere Studies
CORBA	Common Object Request Broker Architecture
DAP	Data Access Protocol
DIMES	Distributed Metadata Server
DOA	U.S. Department of Agriculture
DOC	U.S. Department of Commerce
DOE	U.S. Department of Energy
DODS	Distributed Oceanographic Data System
ECMWF	European Center for Medium-Range Weather Forecasting
EOS	Earth Observing System
EPA	Environmental Protection Agency
ETA	NCEP Eta Model
ERS	European Remote Sensing Satellite
ESG	DOE Earth System Grid
ESML	Earth Science Markup Language
FGDC	Federal Geographic Data Committee
FTP	File Transport Protocol
GCM	General Circulation Model
GCMD	Global Change Master Directory
GDAS	Global Data Assimilation System
GDS	GrADS Data Server
GFDL	Geophysical Fluid Dynamics Laboratory
GIS	Geographical Information System

GOES	Geostationary Operational Environmental Satellite
GrADS	Grid Analysis and Display System
GRIB	GRIdded Binary
GSFC	Goddard Space Flight Center
GSM	Global Spectral Forecast Model
GUI	Graphical User Interface
HDF	Hierarchical Data Format
HDF-EOS	Hierarchical Data Format - EOS
HIRS	High-Resolution Infrared Radiation Sounder
HTML	Hyper Text Markup Language
HTTP	Hypertext transport protocol
IDL	Interactive Display Language
IEEE	Institute of Electrical and Electronics Engineers
IIOp	Internet Inter-Orb Protocol
IPCC	Intergovernmental Panel on Climate Change
JGOFS	Joint Global Ocean Flux Experiment
LAS	Live Access Server
LLNL	Lawrence Livermore National Laboratory
MIT	Massachusetts Institute of Technology
MSU	Microwave Sounding Unit
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCDC	National Climatic Data Center
NCEP	National Centers for Environmental Prediction
NDARS	NOAAPort Data Access and Retrieval System
NERC	Natural Environment Research Council
NetCDF	NETwork Common Data Format Data Access Protocol
NGI	Next Generation Internet
NGM	NCEP Nested Grid Model
NOAA	National Oceanic and Atmospheric Administration
NOPP	National Oceanographic Partnership Program
NSDL	National Science Digital Library
NSF	National Science Foundation
NVODS	NOAA Virtual Ocean Data System
NWP	Numerical Weather Prediction
NWS	National Weather Service
OPeNDAP	Open source Project for a Network Data Access Protocol
PCM	NCAR Parallel Climate Model
PCMDI	Program for Climate Model Diagnosis and Intercomparison
PMEL	Pacific Marine Environmental Laboratory
RUC	NCEP Rapid Update Cycle Model
SSI	Spectral Statistical Interpolation
SSM/I	Special Sensor Microwave Instrument
SST	Sea Surface Temperature
TCP/IP	Transmission Control Protocol/Internet Protocol

THREDDS	Thematic Real-time Environmental Data Distributed Services
TIROS	Television Infrared Observation Satellite
TOVS	TIROS Operational Vertical Sounder
UCAR	University Corporation for Atmospheric Research
URI	University of Rhode Island
URL	Uniform Resource Locator
USWRP	U.S. Weather Research Program
WRF	Weather Research Model
WWW	World Wide Web
XML	Extensible Markup Language

## 1.0 Background

A major transition in weather and climate prediction is now occurring, one in which real-time and retrospective climate and weather prediction is spreading from a handful of national centers to dozens of groups across the country. This growth of global and regional scale Numerical Weather Prediction (NWP) and General Circulation Model (GCM) development is now possible due to the availability of:

- multiprocessor workstations;
- regional scale models that run on these workstations (e.g., MM5);
- analysis and forecast grids from modeling centers; and
- high temporal resolution forcing data from historical and future climate scenarios that run with climate system models.

With the availability of these opportunities, in conjunction with the common data access framework, as detailed within this program plan, and easy access to multi-institutional, multi-discipline coupled model results (land, atmospheric, ocean & sea ice, carbon cycle, etc.), carefully designed systematic model intercomparisons are greatly facilitated.

The National Center for Atmospheric Research (NCAR), the Climate Diagnostics Center (CDC), and the National Climatic Data Center (NCDC) provide observations and some gridded data generated by GFDL, NCEP, and a limited number of other institutions. However, these data are very limited in both spatial and temporal resolution. Further, each of the various data sets are provided in various formats. Thus, there is no routine source of historical NWP data from NCEP, or GCM data from GFDL, with no framework process to intercompare these data sets. Rapid development in communications technologies over the last decade has been essential to provide significant gains in our ability to access our national data archives. These technologies have been expected to enable collaborative research between institutions to rapidly grow. However, the difficulties in achieving accessibility and interoperability in these national archives have delayed the realization of these potential gains.

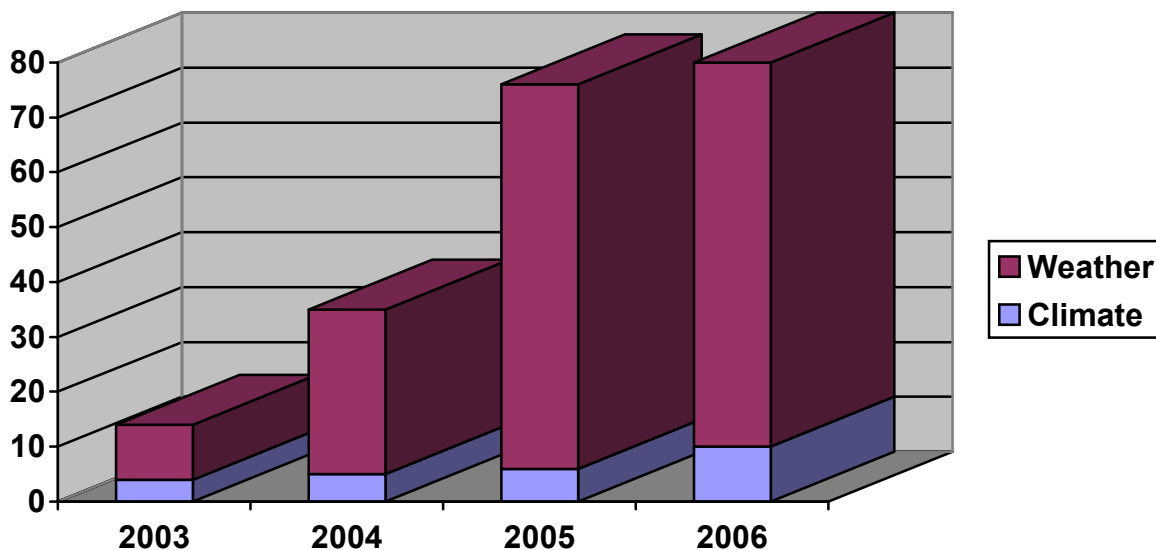
### **A New Framework: Distributed Access and Computing**

A new paradigm for sharing data among users is evolving. It takes advantage of the Internet and the relatively inexpensive computer hardware. In this new model, data providers put their data onto a computer on the Internet. Software running on the computer allows outside users to see not only the data local to that computer but also data on other computers running this same software. In this new scheme the data format is not very important. The software can convert the data into many different commonly used data formats. The only requirement is that the software knows the data format. The NOAA Operational Model Archive and Distribution System (NOMADS) project has adopted this new distributed paradigm for sharing data. In addition, it has collaborative efforts with other distributed software systems. It is becoming clear that this new paradigm will become the dominant model for sharing data in the near future. NOAA has a unique opportunity to be at the forefront of this new paradigm in sharing data. It can easily be envisioned that the distributed paradigm scope could encompass the fabric of scientific progress in this Nation. The philosophy of the "Grid" uses this paradigm. The current state of data

autonomy is undesirable for several reasons. It requires every group and project in the organization responsible for data to design and implement their own means for making that data available to others, including awareness of the best practices for metadata representation for discovery and use, knowledge of how to make the data useful to a larger set of current and future uses than the specific project that generates the data, and resources for providing all the data services that are needed for efficient access to the data. Data autonomy and limited resources lead to lack of awareness of beneficial connections among the data collections, and no means to readily determine how to access data from other groups in the organization, despite the fact that some organizational infrastructure is common to a wide range of data products and services.

NOAA can foster interoperability by integrating demonstrated and working systems, relying on local decisions about systems that have evolved to successfully occupy a data niche, rather than imposing from the top standards that may be inappropriate. Loosely combining legacy systems while developing new ways to support data access to valuable data assets would permit NOAA to work on the cutting edge of distributed data systems.

The scope and growing volume of data providers at NOAA is very large (Fig. 1). Output from computer models, satellite imagery, real-time and retrospective radar and observations, and GIS-based data sets are just a few examples of the variety of data available at NOAA, and potentially available under the NOMADS framework. Weather model volume significantly increases over the next 3 years, however not all forecast files will be permanently archived. The leveling off of archive growth seen in Fig. 1. represents the cycling of older forecast data sets no longer required with new data. It is expected that forecasts will be saved for 5 years and those not selected for long term archive, will be removed.



**Figure 1. Climate and Weather Data Archive Growth (Pbyte)**

#### **Purpose and Scope**

This Program Plan will be updated with Appendices, while the core data management vision

within the body of the plan will remain mostly unchanged. This document provides:

- a description of the effort;
- how the system is used (by a data user or a data provider);
- roles and responsibilities for core participants with a focus on NCDC milestones;
- provides the initial data and system requirements;
- archive requirements for NCDC;
- a project schedule;
- minimum resource requirements for core participants,
- a data management vision for use by NOAA IT management in supporting NOMADS.

### **Program Goals and Objectives- An Overview**

A partnership among data centers is required to establish the NOMADS. This program plan is the initial step in this potential long-term collaboration. NCEP has expertise in running its forecast systems, NCAR and NCDC have great experience in running a data distribution center, while the

PMEL, LLNL and COLA, will provide a technology transfer component and a Web Portal and data browsing system. For NWP, NCEP will provide the bulk of the raw data necessary to populate the joint archive. Eventually, both NCEP and GFDL will populate a climate archive using the same technologies for the initial NOMADS implementation. A partnership worked extremely well for the NCEP/NCAR Global Reanalysis project and NOMADS is an extension of that collaboration. This partnership or division of labor in NOMADS, will play to each organization's strengths. Under this project, NCEP would also develop NWP data requirements to fulfill NOMADS objectives and provide that data to NCDC and NCAR using common access and dissemination methods as described in this document.

Not unlike the explosive growth and sharing of resources as seen on the Internet over the last decade (for textual based data), the NOMADS project enjoys the cooperation of many differing institutions and laboratories and takes advantage of many man-years of software development across NOAA and elsewhere. NOMADS is a logical extension of the capabilities of the Internet for a science-based data and product information exchange for the purpose of climate and weather model improvements and verification.

### **NOMADS Program Goals**

In NOMADS, no one institution carries the weight of data delivery since data are distributed across the network, and served by the institutions that developed the data. The responsibility for documentation falls on the data generator; with the NOMADS Advisory Teams ensuring overall quality and systems standards, and to determine which NOMADS data are required for long-term storage. Further, NOMADS in no way precludes the need for national centers to maintain and support long-term archives. In fact, NOMADS and secure data archives are mutually supportive and necessary for long-term research. The primary science goal of the NOMADS framework is that it enables a feedback mechanism to tie Government and university research directly back to the NOAA operational communities, numerical weather prediction quality control and diagnostics processes at NCEP, and climate model evaluation and inter-comparisons from around the world.

The goals of NOMADS are to:

- improve access to NWP and GCM models and provide the observational and data assimilation products for Regional model initialization and forecast verification,
- improve operational weather forecasts,
- develop linkages between the research and operational modeling communities and foster collaborations between the climate and weather modeling communities,
- promote product development and collaborations within the geo-science communities (ocean, weather, and climate) to study multiple earth systems using collections of distributed data under a sustainable system architecture, and
- provide a long-term framework for systematic approaches to climate change detection efforts, climate and weather model evaluation, impacts studies, and other process studies.

Within the NOAA community, the primary objective of NOMADS is to preserve and provide retrospective access to NCEP NWP and GFDL General Circulation Models (GCM) and simulations and data and make these data available to government and University researchers. NOMADS will also provide access to operational NWP guidance products to National Weather Service (NWS) forecasters for Case-Study training. Users will be able to retrieve climate and numerical models in a timely and format independent manner from centrally configured servers.

The NOMADS infrastructure was originally proposed in the U.S. Weather Research Program (USWRP) Implementation Plan for Research in Quantitative Precipitation Forecasting and Data Assimilation to "redeem practical value of research findings and facilitate their transfer into operations."

Since most weather model data are stored in the GRIB and BUFR data formats and most climate model data are stored in the NetCDF data format, NOMADS will provide for the comparison between these two formats. Converting between these two data formats has been a very difficult task in the past. The NOMADS software allows users to work in whatever data format is the most advantageous to them. The combination of quality control routines, and independent data format will, for the first time, provide users with a seamless interface to models and associated data. Climate modelers could access weather models, and weather modelers could access climate models. Never before has this capability existed outside a given institution.

NOMADS allows for growth of data sets and for changing data formats. If the current list of applicable formats no longer applies or is no longer in use, the NOMADS framework allows the XML to conform to the new data "standard." Although XML is supposed to be a generic and general language, in real life it requires the prior agreement on a vocabulary by all parties. Thus, full coordination is required for a successful inter-operability of systems for data sharing. This coordination already exists among the NOMADS collaborators as is evident by the sheer number of collaborating institutions contributing their expertise and collections under the NOMADS framework. The NOMADS distributed client-server architecture has the potential to bridge the gaps of inter-operability between systems and thus provide users with a metadata of geo-sciences information, visualization and analysis tools, and research opportunities- potentially reaching

beyond the physical sciences.

### **NOMADS Operating Principles**

- NOMADS is an agreement between agencies, who participate, to have common data and observation distribution software, using a format independent and description methodology, and for a unified documentation and organizational framework for data distribution.
- NOMADS provides a forum to plan and organize these structures for use by university, federal agencies and organizations as providers (servers of the data) or as clients (the users of the data).
- NOMADS is a framework for these agencies or institutions to obtain support for the dissemination of their data sets and to make use of these data sets. NOMADS participants will provide their own resources to accomplish this. A resource means the hardware and the programming staff to execute their plans. Clients should invest in the training it takes to use a web page client, like ftp2u, or client software like GrADS, IDL, MatLab, Ferret, Live Access Server, idv, etc.
- NOMADS will integrate Grid based technologies and the Open Source Python language, to allow easy access to large data sets and time series of multiple files (e.g., 5 years of u/v from the Eta model) in a modular programming language. Through choices of user services- GDS, LAS, OPeNDAP, HTTP, ESG, and CEOS-Grid, these choices will allow for model, observational and other (i.e., satellite data in HDF format, and OpenGIS data structures), data using LLNL developed Climate Data Analysis Tools. Under Python, and the modular Open
- Source design, other discipline specific routines would become available to the NOMADS user (e.g., high energy plasma physics statistics routines would be “plug and play” for the NOMADS user.

### **The Geosciences Community**

In a larger context NOMADS can lay the groundwork to achieve interoperability between the geosciences community, allowing each participant full control of its own goals and objectives. NOMADS is a working example that fulfills the goals as outlined by the CCRI, Chapter 12 “Grand Challenges in Modeling, Observations, and Information Systems” by providing format independent access to heterogeneous data across multiple disciplines and institutions.

### **Roles and Responsibilities**

This document will outline roles and responsibilities for participating NOMADS laboratories and data centers. This is a living document, and as technology advances, or users needs change, so too will specific roles and responsibilities. There are five primary or core NOAA NOMADS partners: CDC, GFDL, NCDC, NCEP and PMEL. The exact duties within each of the below listed participants are not spelled out explicitly. Rather, the overarching role each Center contributes with respect to their participation under NOMADS is described. Upon formal review, each participant can modify these responsibilities. This will be performed through NOMADS workshops and informal meetings.

External partners are a critical link in the NOMADS distributed data philosophy. Without multi-Agency, and International participation, the goal for multidiscipline, interoperable data services

breaks down. Therefore, roles and expected responsibilities at participating Agencies, Programs (i.e., Earth Systems Grid), and International partners, external to NOAA, are also included.

Specific NCDC roles and responsibilities are provided along with milestones and deliverables to address NCDC specific staff and resources requirements. It also serves as a catalyst for an expanded NOMADS activity within NCDC operations (i.e., Access and Archive Branch involvement).

### **NCDC Participation**

NCDC participation will include 1) the ingest and archive of the operational NCEP gridded model output fields and 2) the ingest and archive of the NCEP Global Data Assimilation Analysis System (GDAS), both for NOMADS access. Initially, mode output fields will include the gridded models available over NCDC's NOAAPort System accounting for approximately 1TB/year. However, the NOAAPort broadcast has a limited number of output parameters, and at a lower resolution currently being generated at NCEP. NOMADS will initially provide these NOAAPort data for NOMADS access but plans are underway to the ingest and archive (non-permanent) the full suite of NCEP gridded output. Since these data volumes are very high, and the need for forecast files in the long term cannot be made, these data will be rotated after a period of time determined by the NOMADS Steering and Science teams (approximately 2-5 years). The data volume expected for these highly requested data approach 70TB/yr.

Climate models generated at GFDL (R15, and R30), are currently being served at GFDL under NOMADS, and long-term stewardship of these models may take place at NCDC if so requested by the Science team and required by the Data Archive Board (DAB). For a detailed description of NCDC's activities see Section XXXX.

### **NCDC Deliverables**

NCDC participation will focus on the ingest and archive of the entire NOAAPort data stream including NCEP AWIPS model output grids, and other products from NWS sources. NOMADS provides 30-day on-line access (via DODS/OPENDAP in NetCDF format) to these AWIPS/NOAAPort operational models for use by NWS forecasters and by the Cooperative Program for Operational Meteorology, Education and Training (COMET) program. Older data will be migrated to NCDC archive systems and development of Web-based access will be developed in FY02. GFDL will provide limited GCM data to NCDC for archive and access while NCAR will continue to provide access to NCEP reanalysis data sets under this infrastructure.

#### **FY03**

1. Develop the NOMADS Program Plan with a Users Guide and to develop a Federation of participants. This Federation will provide attribution for participating institutions, and provide background and detailed information about NOMADS, its members, how to become a member, listing of quarterly reports including technical specifications. NCDC will lead and implement a NOMADS Web Page for this purpose.
2. Implement a NOMADS Steering Committee.

3. Conduct a Science Planning Workshop.
4. NCDC must include it's Data Access Branch (DAB) staff with NOMADS collaborators to ensure NCDC operations are actively involved in this rapidly growing project, and collaboration in terms of technical data distribution efforts NOAA-wide.
5. NCDC to develop science-based partnerships with GFDL, NCAR and elsewhere. NOMADS is data distribution and data access project; whereas science based activities must begin either separately or in concert with the NOMADS. In order to achieve these new collaborations, NCDC will advance the use of these technologies within its Scientific Services Division.
6. NOMADS will work toward the implementation of Grid technologies within the Earth System Grid (ESG), and the CEOS DataGrid. This will advance NCDC's understanding of grid technologies including a functionality called the Hierarchical Resource Management (HRM) that would interface NOMADS spinning disks with HPSS-based mass store archives. HPSS is currently in use within the NCDC main archive.

### **NCDC NOMADS Disk to Mass Storage Interface**

NCDC will develop a NOMADS spinning disk to Mass Store capability. The NCDC NOMADS Team, are currently installing the hardware and software for a minimal NOMADS service. The goal for FY03 is to initiate the HDSS (HAS) interface work to reach NCDC's archive from NOMADS spinning disk servers. It is expected that this activity will extend into FY04.

The NOMADS collaboration with the DOE's Earth System Grid (ESG), and the CEOS-DataGrid, will play an important role in a technology transfer process. These inter-Agency programs are developing the systems and security processes to allow Internet to mass storage access. NOMADS will leverage these efforts and adapt them to function within NCDC. Archive data sets at NCDC use HSM, indexes, and relational databases for HDSS retrieval. The new grid-based services will provide for access in a secure environment.

Each participating data provider with archive services (HDSS etc.) will have to develop its own disk to mass storage capability independently since the configuration at each center will be slightly different. There will be common approaches for access, and as this capability develops each center can document specifics as to implementation to assist new data providers. Currently only NCDC and NCAR are investigating this capability.

NCDC Archive Branch- TBD

NCDC Data Access Branch - TBD

### **The NOMADS Framework**

The implementation of the NOMADS framework by a participant implies acceptance of responsibilities including dedicated support for data set maintenance, documentation and software programming and local web services and portals. A commitment to support the NOMADS hardware and software and common data access methodologies are integral parts of

the program. In addition to the core participation of NOAA laboratories, centers and offices, NOMADS participants may also be a university, federal agencies, and organizations who act according to participant guidelines. NOAA participants are servers of data and are partially responsible for the support of their own NOMADS resources, in terms of contributing unique dataset knowledge to document necessary tasks.

Users of the data on the servers are known as clients and are NOAA laboratories, centers and offices as well as university, federal agencies and other external organizations that extend to the public arena. NOMADS fills a need and satisfies a NOAA responsibility for archive and real-time access to data sets including the ability to dissect and download user selected products. NOMADS is a common sense approach to satisfy a NOAA responsibility to public and federal organizations to disseminate and archive a plethora of datasets in a cost effective way.

The NOMADS framework is a distributed data access and archive infrastructure and promotes the combining of data sets between distant participants using open and common server software and methodologies. Users effectively access model and observational data and products in a flexible and efficient manner from archives or in real-time through existing Internet infrastructure.

NOMADS will implement an open, and distributed data service through the agreement and participation of core NOAA, and non-NOAA data providers. The data format neutral “glue” that holds the effort together is OPeNDAP (formally known as DODS/OPeNDAP). OPeNDAP is a software framework, and consortium that is a viable project with project resources. NOMADS uses this framework and use of these data conventions (cf, NetCDF, or any OPeNDAP enabled client) greatly advances (but does not ensure) for cross discipline interoperability (including ocean and in-situ) between the climate and ocean communities. The team agreed that the current level of interoperability was sufficient for the short-term, but that steps need to be taken to ensure for use of new data forms in the future (e.g., satellite and radar).

OPeNDAP is a binary-level protocol designed for the transport of scientific data subsets over the Internet. Through OPeNDAP applications can open and read subsets from remote data sets. OPeNDAP is central to the NOMADS distributed philosophy for successfully distributing classes of data from oceanography, meteorology, climate research and other geosciences data and thereby providing a solution for interoperability problems. NOMADS success depends on supporting software development technology transfer. In this case, to receive technology transfer and support from NOAA/PMEL and OPeNDAP where the National Ocean Partnership Program (NPP) has emerged as a respected standard for environmental data access and augment the NOMADS program.

To learn and adopt the OPeNDAP protocol, UCAR’s Unidata Program has assumed a national role for the understanding and advancement of OPeNDAP. See <http://www.unidata.ucar.edu/packages/DODS/OPeNDAP/>.

## **NOMADS Teams and Working Groups**

Three NOMADS Teams have already been established. These include:

1. The NOMADS Steering Group

The NOMADS Steering Group are comprised of PI's from each core participant and others as selected by the Group as a whole. The group, in coordination with the Science Team, determines archive requirements for selected data sets; and helps to direct the long-term vision of NOMADS. The Group will solidify understandings among the primary lab's and data centers and move NOMADS from this grass roots effort into a mainstream NOAA level, NOAA supported (via line office) program. It is proposed that the members be selected from the core collaborator's: GFDL, LLNL, NCAR, NCDC, NCEP, and PMEL. The steering group will report to NOAA. The Group should meet one per year. See Appendix A. for a current listing of the Steering Group, the Advisory Team and their charter. The Advisory Team members will grow as recommended by the Steering Group.

2. The NOMADS Science Team

The NOMADS Science Team will direct the requirements process for on-line access and long-term stewardship and make recommendations to the NOAA Archive Board. They will help define climate and weather tools necessary for systematic approaches to climate model evaluation; and define and develop tools for quality control of various models and data. Most of these efforts can be accomplished by email and informal discussions. The Science team will be responsible for long-term science planning and associated formal workshops called by the Steering Group.

3. The NOMADS Technical Team

In a partnership with the Science Team and Steering Group, the NOMADS Technical team will oversee the implementation of necessary systems and access configuration decisions. The Team is attempting to meet twice a year. The first meeting was at the NOAA Science Center (NCEP), June 28, 2000 with 4 workshops since that time. Studies and benchmarks were performed and the NOMADS Steering Group agreed to use the Open source Project for a Network Data Access Protocol (OPeNDAP, see section XXX below). Through a series of formal and informal meetings and agreements, the NOMADS Technical Team will develop and recommend follow-on enhancements to "NOMADS Version 1.0" to the NOMADS Steering Group for coordination and provide implementation dates for core members. The Technical team will also lead the Web Portal effort for the development of a common user interface for use inside and outside NOMADS. The Technical Team(s) will focus on more immediate tasks. There could be more than one of these technical teams. They would have to communicate with each other much more frequently than the Steering Group (both among the team members and between teams). These technical teams are solving real problems and can meet in an ad-hoc fashion using the Steering Group and Science Team for guidance.

## **Science Based Planning**

It is well known that a major obstacle for climate change detection and attribution studies is the lack of systematic long-term approaches to model evaluation. The NOMADS software

infrastructure is a framework that promotes coordinated approaches to model evaluation. With NOMADS and easy access to multi-institutional, multi-discipline coupled model results (land, atmospheric, ocean & sea ice, carbon cycle, etc.), carefully designed systematic model intercomparisons are greatly facilitated. The NOMADS required tools, data availability, and long-term archive, needs to be defined in terms of science-based requirements and planning.

A workshop of stakeholders is proposed to define a scientific baseline climate model and observational data requirements document. It is required to define data archive requirements under NOMADS for climate, ocean, and weather process studies from global to regional, seasonal to decadal. Climate change detection methods for inclusion in NOMADS will be identified. These will include (but not limited to), optimal detection regression techniques (e.g., fixed pattern, space-time, space-frequency), ensemble calculations, and other process studies such as pattern correlation and dimension reduction. In addition, data requirements for various impact studies need to be identified.

A series of workshops are required to focus on the needs of these scientists, in terms of available software tools, to coordinate systematic approaches to these studies across disciplines and institutions using the NOMADS compatible clients (e.g., DODS/OPeNDAP, LAS/Ferret, GrADS, IDV, IDL, CDAT, Matlab, and Python based systems). These data analysis tools may include:

- Eigenvalue techniques and matrix-oriented operations; analysis of the joint spatial and temporal variability of scalar or vector fields like geopotential height, precipitation or temperature over a wide area to identify the main modes of variability embedded in a temporally and spatially variable data set, which is often performed by means of Empirical Orthogonal Function (EOF) approach, based on the Karhunen-Lo  ve decomposition of the joint temporal and spatial variability of the fields.
- Similarly, during the analysis of coupled variability of geophysical data sets, eigenvalue techniques like Canonical Correlation Analysis (CCA) or the Singular Value Decomposition (SVD) of the covariance matrices are potential required tools.

These workshops will further the preliminary NOMADS vision, and help to develop additional partnerships and funding approaches. They will also serve to further coordinate with existing collaborative projects, including model coding standards efforts such as the US based Earth System Modeling Framework (ESMF) and the European based PRISM project. Both of these programs are model coding frameworks that modeling centers are now adopting. These programs will provide for a modular coding design to promote model enhancements across institutions.

### **Web Portal Development**

CDC, GFDL, and others, specifically the IRI at Columbia, and PCMDI at LLNL have extensive experience with distributed model and in-situ Web Portal services. The NOMADS group needs to better leverage from their experiences. Currently, the back-end services (GrADS, Live Access Server, DODS/OPeNDAP) are well defined, with software development groups leading the way.

NOMADS will not substantially modify these groups' efforts. What is needed, and where NOMADS can be most effective during the next year is the development of easy to understand, front-end GUI / Web interfaces to NOMADS services. Currently each participation lab or Center is developing it's own Web interfaces. The interface should be modular so that a participating Center can pick and choose which NOMADS services they desire for internal use or public access.

NOMADS members are currently developing a common web portal across collaborating institutions. This effort is underway between NOAA, NASA, DOE, NCAR, and the British Atmospheric Data Center (BADC) and the European "e-Science" program. A second meeting is planned for early 2003 that will focus on:

- General agreement on what should be included in a common web portal design and a summary of these requirements.
- A list of tasks that need to be performed to build the common web portal and a corresponding list of what each institution can contribute to the development of these tasks. The discussions will focus on the following areas:
  - Who are the intended users and what basic tasks are needed from a web interface?
  - Survey of current web interfaces and their strengths and weaknesses. (U.S.-GFDL and European-British Atmospheric Data Centre).
  - What each institution thinks their web portal would look like.
  - Define the requirements for a customizable web interface.

NOMADS participants may incorporate, in an open technology transfer component, this Web portal development into their operations as necessary. Requests for these developments should be made through a NOMADS Steering Group member (see Appendix XXXX).

### **NOMADS System Architecture: Initial Operating Capability (IOC) for v.1**

NOMADS core sites will serve their data both through OPeNDAP (included in the GrADS-Data Server) and through LAS. Each site may have a unique LAS back end that incorporates their in-house tools (CDAT, NCL, IDL, MatLab, Ferret, GrADS, etc.). The NOMADS team will build a unified Web entry point to all of the data using a LAS "sister server" configuration, where OPeNDAP makes comparisons at the binary level possible. A CDAT back-end in the unified server would provide a mature model inter-comparison metrics framework. A range of OPeNDAP-enabled desktop tools would be available for detailed, custom analysis: Matlab, IDL, GrADS, Ferret, Excel. Our goal is to migrate the entire framework to work within the GRID (Earth System Grid, CEOS-Grid).

NOMADS is a user of the OPeNDAP framework. OPeNDAP is a stand-a-lone server. The DAP protocol (OPeNDAP's core) has no dependencies to the HTTP library or any other communications library. DOE's Earth System Grid (ESG) has been working on a new Point-to-Point Transfer (PPT) communication library for OPeNDAP. This lightweight library allows easy communications with OPeNDAP servers so data can be transferred from OPeNDAP servers to

GridFTP servers. GridFTP uses Globus Grid technology to transfer large amounts of data securely and quickly. These technologies will be incorporated into NOMADS versions 2 and beyond as they are developed.

### **Templates for Data Users and Providers**

The following is provided to acquaint data provider and data users with data access and services philosophy under NOMADS.

#### **A NOMADS Data Provider**

A NOMADS participant, in this case the agency or project disseminating the data set, controls the integrity and documentation of datasets. The display and scientific manipulation of data will be accomplished by the client (data receiver), who elect to use one of many compatible public or commercial display packages. A NOMADS server has the ability to dissect data sets and recompose fields and make calculations as desired by the client. Web based ftp can also “slice and dice” data sets according to users needs as part of the NOMADS framework.

The NOMADS collaborators agree that the continuing success of NOMADS depends upon some level of participant resources to support to 1) a steering committee; 2) a NOMADS software control and NOMADS Web Portal team; and 3) support for OPeNDAP.

To become a data provider under the NOMADS framework, a data center or researcher must

- 1) Enable their data for traditional or distributed access. For distributed access, the adoption of the OPeNDAP protocol is required.
- 2) A NOMADS data provider shall document and maintain, in plain language, all data and metadata for all required models and data. The Extensible Markup Language (XML) is used for this purpose. If the NOMADS Science Team has recommended the data for long-term storage at NCDC, documentation of data and metadata shall be provided under the FGDC documentation standards.

#### **Conventional Data Access**

In addition to the OPeNDAP server, regular ftp services may be provided for access to data sets not necessarily applicable under the distributed framework, i.e., entire data sets (although entire data sets may be obtained under NOAMDS), or users desiring to use standard download services. One existing service is already provided by the Real Time NOMADS (RT-NOMADS) project at NCEP to serve the operational analysis ready observations and model run history using [ftp2u](#), a web based file transfer protocol (ftp). The ftp2u delivers unaltered data sets as well as user defined “slice and dice” data sets by parameter space, physical space, and time. Users can subset large files of high resolution model results and regroup different data sets to create needed products.

NOMADS will also develop and implement metadata interfaces to existing NOMADS partners serving compatible data sets. This currently includes NCAR’s Community Data Portal, the National Oceanographic Partnership Program’s (NOPP) National Virtual Ocean Data System

(NVOADS); the Department of Energy's Earth System Grid (ESG); and the Thematic Real-time Environmental Data Distributed Services (THREDDS) project being developed through the National Science Foundation and Unidata, the OPeNDAP, the OPeNDAP Data Connector (ODC), the PMEL and LLNL developed LAS/CDAT access and diagnostic package, and the GrADS-DODS/OPeNDAP Server (GDS) developed at the COLA; and the NCEP Real Time NOMADS project called RT-NOMADS. First priority will be to develop NOMADS compatibility with OPeNDAP protocols in the NVOADS and OPeNDAP environments.

### **Data Users under NOMADS**

Currently, GFDL has one of the larger NOMADS sites for climate data. On the GFDL NOMADS server there is model data available from various climate studies, from the seasonal-interannual to decadal and longer time scales. There are more than 700 GB of model data available in more than 4,000 files. From October, 2000 to August 2002, there was more than 480 GB of data downloaded by users. There are about 75 registered users, 80% are from the United States and 20% are international users. Of the 75 users, about half are from private groups and corporations, with about a 25% from universities and about 15% from government users. Most of these users are scientists interested in obtaining GFDL model data.

The users of the GFDL data range from people interested in performing studies of how the climate of Arctic region changes due to changes in the greenhouse gases to the IPCC Data Distribution Center. A typical user finds the model data via "word of mouth". The user then searches the GFDL site, for the data needed. At present, this process is rather difficult and is representative of an area of active software development for NOMADS. The user then downloads the model data via ftp to a local computer where the user performs some analysis using the model data. The model data is stored in a netCDF data format, so the analysis process is normally straightforward if the user has experience using netCDF data.

### **Open source Project for a Network Data Access Protocol: OPeNDAP**

The Open source Project for a Network Data Access Protocol (OPeNDAP- formally known as the Distributed Oceanographic Data System (DODS/OPeNDAP)) began as a joint effort between staff and scientists at the University of Rhode Island, Graduate School of Oceanography and at the Massachusetts Institute of Technology, Department of Earth Atmospheric and Planetary Science. DODS/OPeNDAP is a software framework that simplifies all aspects of scientific data networking, allowing simple access to remote data. Local data can be made accessible to remote locations regardless of local storage format by using DODS/OPeNDAP servers. Existing, familiar data analysis and visualization applications can be transformed into DODS/OPeNDAP clients, i.e., applications able to access remote DODS/OPeNDAP served data. DODS/OPeNDAP provides a protocol for requesting and transporting data across the web. The current DODS/OPeNDAP Data Access Protocol (DAP) uses HTTP to frame the requests and responses. For details on the DODS/OPeNDAP DAP, see DODS/OPeNDAP Data Access Protocol (DAP 2.0) at [www.unidata.ucar.edu/packages/DODS/OPeNDAP/design/dap-rfc-html](http://www.unidata.ucar.edu/packages/DODS/OPeNDAP/design/dap-rfc-html).

The OPeNDAP involves a community of users working together to use, improve, and extend the OPeNDAP protocol and software. The OPeNDAP design principles are based on two considerations:

- data are often most appropriately distributed by the individual or group that has developed them;
- the user will in general like to access data from the application software with which s/he is most familiar.

This has resulted in a highly distributed system that allows users to control the distribution of their own data and the way they access data from remote sites.

### **NOMADS Desktop Services**

For data users, the currently available OPeNDAP servers include data formats for many desktop applications, and the user base for OPeNDAP enabled clients is expanding. NOMADS users have multiple options for access data including:

DODS/OPENDAP geturl	a simple command-line client for testing DODS/OPENDAP datasets
DODS/OPENDAP IDL Command-line client	an IDL tool which provides access to DODS/OPENDAP data in IDL. IDL (a commercial data analysis and visualization package) is required.
DODS/OPENDAP Matlab Command-line client	a Matlab tool which provides access to DODS/OPENDAP data in Matlab. Matlab (a commercial data analysis and visualization package) is required
DODS/OPENDAP Matlab Toolkit (GUI)	a GUI for accessing oceanographic data via DODS/OPENDAP in Matlab (a commercial data analysis and visualization package).
Ferret	a data analysis and visualization package available from NOAA/PMEL that can access data via DODS/OPENDAP
GrADS	a data analysis and visualization package available from COLA that can access data via DODS/OPENDAP
IDV (Integrated Data Viewer)	a Java application for visualizing and analyzing geoscience data
LAS (Live Access Server)	a highly configurable Web server designed to provide flexible access to geo-referenced scientific data.
ncBrowse	is a Java application that provides flexible, interactive graphical displays of data and attributes from a wide range of netCDF data

	file conventions
ncdump	a tool for displaying netCDF files as netCDF CDL
NCO (netCDF Operators)	NCO is a set of command line tools that perform operations (e.g., average or concatenate) on netCDF or HDF files. NCO is not distributed as a DODS/OPENDAP client but can be re-linked to the DODS/OPENDAP libraries.
ncview	a visual browser for gridded data.
CDAT	The Climate Data Analysis Tool developed at LLNL is a flexible open

Table 1. NOMADS Available Clients

### **DODS/OPeNDAP Client Libraries and Other Desktop Applications**

DODS/OPENDAP netCDF client library a DODS/OPENDAP enabled version of the standard netCDF library.

netCDF Java library	the Java implementation of netCDF
---------------------	-----------------------------------

### **Other Desktop Applications**

Generic Web Browser	any generic web browser like Internet Explorer, Mozilla, or Netscape can access OPeNDAP/DODS data in a restricted manner. Since the browsers do not understand the DODS/OPENDAP protocol, the user will need some understanding of the DODS/OPENDAP protocol.
Spreadsheet Applications	(e.g., Excel, StarOffice) any application that can dereference a URL can access a DODS/OPENDAP URL. Since the applications don't understand the DODS/OPENDAP protocol, the user will need some understanding of the DODS/OPENDAP protocol.

### **Table 2. Other applications that can access DODS/OPENDAP Servers**

A data provider can range from a single scientist and a laptop computer, to a major data center. To make data available under the NOMADS framework, a provider must install one of the following NOMADS Available Servers (note: new servers are coming on-line each year):

DODS/OPENDAP netCDF Server	makes netCDF data available via DODS/OPENDAP.
----------------------------	---

DODS/OPENDAP HDF Server	makes HDF 4 data available via DODS/OPENDAP
DODS/OPENDAP HDF5 Server	makes HDF 5 data available via DODS/OPENDAP. (Not yet publicly released)
DODS/OPENDAP RDBS Server	makes RDBMS data (requires JDBC) available via DODS/OPENDAP.
DODS/OPENDAP JGOFS Server	makes JGOFS data available via DODS/OPENDAP
DODS/OPENDAP FreeForm Server	makes data available via DODS/OPENDAP using the FreeForm software package. FreeForm provides a flexible method for specifying data formats thus allowing the DODS/OPENDAP FreeForm server to serve a variety of data formats.
DODS/OPENDAP Aggregation Server	makes data available via DODS/OPENDAP. It's main purpose is to allow the "aggregation of multiple datasets into one DODS/OPENDAP dataset. The DODS/OPENDAP Aggregation Server is currently in Beta release.
GrADS-Data Server	makes data available via DODS/OPENDAP using GrADS. The types of data that can be served include GRIB, BUFR, netCDF, HDF, and GrADS binary.
DODS/OPENDAP Matlab Server	makes Matlab binary data available via DODS/OPENDAP.
DODS/OPENDAP DSP Server	makes DSP data available via DODS/OPENDAP.
<i>DODS/OPENDAP servers not yet publicly available:</i>	
DODS/OPENDAP CEDAR Server	makes CEDAR data available via DODS/OPENDAP
DODS/OPENDAP FITS Server	makes FITS data available via DODS/OPENDAP

**Table 3. DODS/OPENDAP Available Servers**

### Downloading NOMADS Clients and Servers

For more information, or for downloading these services the following Web sites are provided.

#### **DODS/OPENDAP**

For more information on DODS/OPENDAP see:

[www.unidata.ucar.edu/packages/DODS/OPeNDAP](http://www.unidata.ucar.edu/packages/DODS/OPeNDAP).

### **DODS/OPeNDAP Aggregation Server**

The DODS/OPeNDAP AS, is part of the Java/DODS/OPeNDAP Servlet library, allowing physical files to be logically aggregated and served as a single DODS/OPeNDAP dataset. The AS also presents all available datasets in an integrated THREDDS Catalog, and is also a netCDF file server. The AS is currently in beta testing. For more information on the DODS/OPeNDAP AS see [www.unidata.ucar.edu/projects/THREDDS/tech/DODS/OPeNDAPAggServer.html](http://www.unidata.ucar.edu/projects/THREDDS/tech/DODS/OPeNDAPAggServer.html).

### **Live Access Server**

The Live Access Server (LAS) is a highly configurable Web server designed to provide flexible access to geo-referenced scientific data. It can present distributed data sets as a unified virtual database through the use of DODS/OPeNDAP networking. Ferret is the default visualization application used by LAS, though other applications (Matlab, IDL, GrADS) can also be used.

LAS enables the Web user to visualize data with on-the-fly graphics and users can request custom subsets of variables in a choice of file formats. LAS can access background reference material about the data (metadata) and compare (difference) variables from distributed locations.

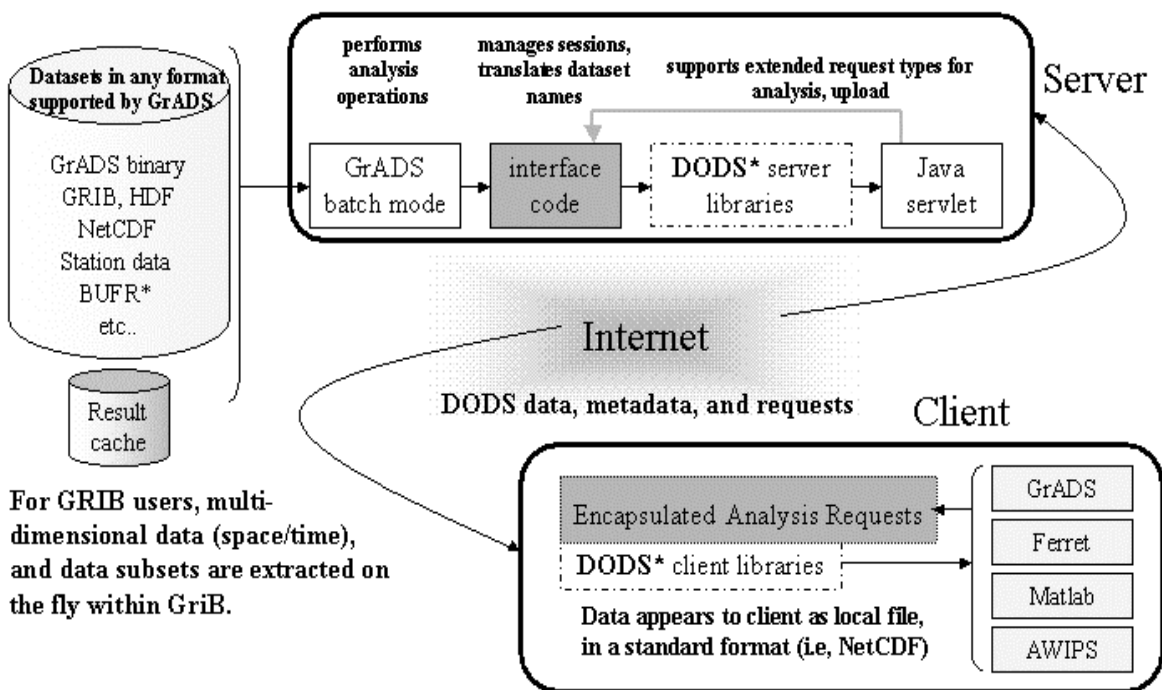
LAS enable the data provider to unify access to multiple types of data in a single interface and create thematic data servers from distributed data sources. LAS also offers derived product generation "on-the-fly", and has the capability to remedy metadata inadequacies, e.g., poorly self-describing data. LAS also offer the ability to produce unique products, e.g. visualization styles specialized for the data, for scientific exploration. For more information on LAS see [http://ferretwrc.noaa.gov/Ferret/LAS/ferret\\_LAS.html](http://ferretwrc.noaa.gov/Ferret/LAS/ferret_LAS.html).

### **GrADS-Data Server**

The Grid Analysis and Display System (GrADS) Data Server (GDS) (<ftp://grads.iges.org/pub/gadev/gds7.doc>) is a data server that provides sub-setting and analysis services across the Internet. These services can be provided for any GrADS-readable dataset. The sub-setting capability allows users to retrieve a specified temporal and/or spatial sub-domain from a large dataset, eliminating the need to download everything simply to access a small relevant portion of a dataset. The analysis capability allows users to retrieve the results of an operation applied to one or more datasets on the server. Examples of analysis operations include basic math functions, averages, smoothing, differencing, correlation, and regression. The GDS supports operations that can be expressed in a single GrADS expression. Figure 1 provides an example of the processes initiated during a typical GDS user request under the GDS NOMADS framework.

There is currently an extensive GDS user base using GrADS as its primary data manipulation client. It is expected that many users of NWP and model input data will access NWP data via NOMADS using GrADS. For further information on GDS see [www.iges.org/grads/gds](http://www.iges.org/grads/gds).

GDS allows for variable comparison. For example, a GDS running at NCAR (<http://motherlode.ucar.edu:9090>) is distributing a set of ensemble members from the "Climate of the 20th Century" runs of the COLA atmospheric general circulation model.



**Figure 2. Example path of a GDS user request**

One can easily compare the relative humidity "rh" from the first two datasets, namely "C20C\_A" and "C20C\_B". If one wants to find a global time-average of the difference at the 1000 mb level in 1960 GrADS can be used as the client to open the following URL as follows

```
ga-> sdfopen
http://motherlode.ucar.edu:9090/DODS/OPeNDAP/\_expr\_{C20C\_A,C20C\_B}{ave\(\(rh.1-rh.2\),time=1jan1960,time=1dec1960\)} 0:360,-90:90,1000:1000,1nov1976:1nov1976}
ga-> display result
```

The analysis results are returned in the variable "result" in the opened dataset. Note that the world coordinate boundaries specified in the third set of curly braces fix the time to 1Nov1976. This can be set to any arbitrary time because the time dimension specification is overridden by the GrADS expression that tells the server to average over the period from January 1960 to December 1960.

In order to facilitate the use of these various servers, a library of GrADS scripts are being developed.

## Climate Data Analysis Tool

Under the NOMADS framework, and collaboration between PCMDI, GFDL, and PMEL, the LLNL has developed a new web access capability that merged the LAS with the CDAT (Williams, et al., 2002) suite of access and quality control programs. The URL to view CDAT-LAS is <http://esg.llnl.gov/las>. Currently, CDAT-LAS is serving up AMIP, CMIP, and NCAR's Parallel Climate Model (PCM) data sets. CMIP and PCM data sets are restricted and only accessed with the proper user name and password. The AMIP data sets are unrestricted and can be accessed by the general public. For more information of the PCMDI effort see [www-pcmdi.llnl.gov](http://www-pcmdi.llnl.gov).

The open nature of the CDAT system will permit any member of the climate community to contribute to the system on an equal footing with the members of PCMDI. With this philosophy, the general goal is to develop a consistent and flexible tool for everyone. NCDC will implement this package during 2003. The collaborations under NOMADS, such as THREDDS and PCMDI with other on-line packages (GDS, LAS, DODS/OPENDAP), are to increase collaboration among climate and weather research scientists. NOMADS will also act as the technology that will allow NOAA to collaborate under the Earth System Grid (ESG) project.

CDAT is:

- portable open source software (free)
- incorporates modules
- exceptions and error-handling
- dynamic typing (for very fast prototyping)
- supports classes; very clear syntax
- extensible in C or C++ and other languages (i.e., FORTRAN)
- access local or remote database servers containing data files in various data file formats
- data extraction, grid transformation, and computation support
- quick and easy way to browse through terabytes of data
- 

For more information on CDAT or ESG see <http://esg.llnl.gov/cdat>.

## Data Cataloging

### The Thematic Real-time Environmental Data Distributed Services Project

The Thematic Real-time Environmental Data Distributed Services (THREDDS) project is a system to make it possible for educators and researchers to publish, locate, analyze, and visualize a wide variety of environmental data in both their classrooms and laboratories. Just as the World Wide Web and digital-library technologies have simplified the process of publishing and accessing multimedia documents, THREDDS will provide needed infrastructure for publishing and accessing scientific data in a similarly convenient fashion.

THREDDDS will establish both an organizational infrastructure and a software infrastructure. A team of data providers, software tool developers, and metadata experts will work together to develop a software framework that allows users to publish, find, analyze, and display data residing on remote servers.

The THREDDDS software framework, based on a concept of publishable data inventories and catalogs, will tie together a set of technologies already in use in existing, extensive collections of environmental data: client/server data-access protocols from the University of Rhode Island and the University of Wisconsin-Madison; Unidata's real-time Internet Data Distribution system; the discovery system at the Digital Library for Earth System Education (DLESE); and an extensive set of client visualization tools. For more information regarding the THREDDDS see [www.unidata.ucar.edu/projects/THREDDDS](http://www.unidata.ucar.edu/projects/THREDDDS).

The newly established National Science Digital Library (NSDL) will focus on "Womb to Tomb" education via the Internet, expand data distribution and sharing, aggregation, and cataloging-across many sciences (biologic, physical, math, etc.). NSDL has a direct link to THREDDDS, and THREDDDS to NOMADS since NOMADS will function as one of the THREDDDS data sources.

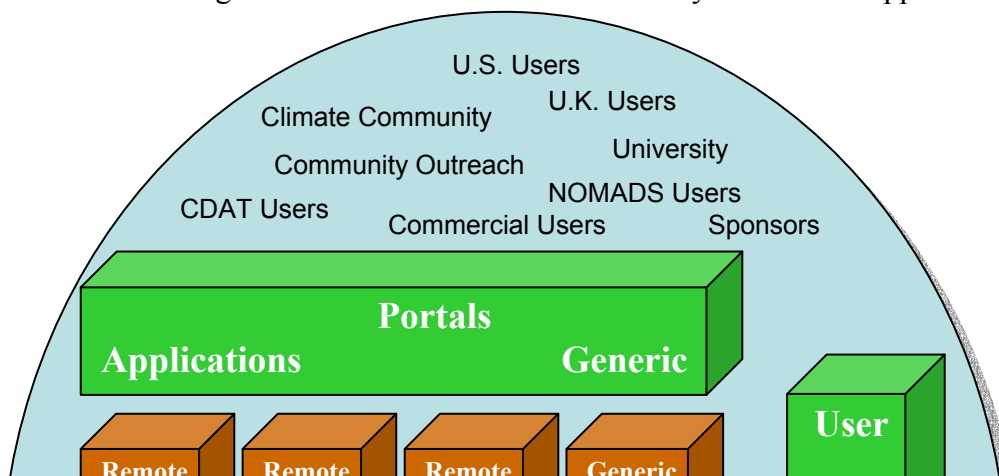
### Distributed Metadata Server

Under the NOMADS and THREDDDS partnership, various database search engines are being explored to locate the many distributed collections of data across the Internet. One such effort is being developed by the George Mason University and is called the Distributed Metadata Server (DIMES) (Yang et al., 2002). In contrast to most other standard metadata systems, DIMES employs a flexible metadata structure, linked data providers (nodes), and supports a wide variety of metadata forms with a minimum of semantic rules. DIMES also provides a software framework to search and browse the metadata. DIMES has been integrated with GDS to create a scientific data information "super-server" to support both data and metadata access consistently. One such example is running at GMU. For more information see: <http://spring.scs.gmu.edu:8099/servlet/SiesipDataTree>.

### GRID Services under NOMADS

Grid technologies and middleware are in the process of transforming the way science will be conducted around the globe. There are currently over 100 national and international grid projects involving many different disciplines, organizations, collaborators, facilities and instruments in literally dozens of countries.

Figure XXX. An Integrated Grid Architecture to Enable Dynamic Grid Applications



Despite the extensive and on-going development of grids, tools, applications, and middleware, the long term viability and persistence of these technologies is uncertain. And, if these technologies and middleware cannot be made persistent, this will not only threaten the future of science and research, but also may preclude and severely limit science itself because scientists cannot afford to base their research on technologies which may not be supported or available in 2-3 years. Toward that end, NOMADS relies on high level grid practices and management oversight currently being performed by the international Middleware And Grid Infrastructure Coordination Committee (MAGIC) group. NOAA is represented on this committee. See [http://www.nsf-middleware.org/MAGIC/Blueprint\\_Overview.htm](http://www.nsf-middleware.org/MAGIC/Blueprint_Overview.htm) for more information.

Beyond the national goal of grid interoperability, NOMADS intends to incorporate advanced data access processes through these partnerships. Open source Grid-based technologies will be used in NOMADS to perform access and Grid services such as the Hierarchical Storage Resource Manager (HRM) may be implemented to interface with mass storage devices, (HPSS and other mainframe archive systems) for retrospective access to institution specific data. Other near real-time Grid services such as “Grid-FTP” will allow for high volume point-to-point data transfers between Centers and Laboratories.

NOMADS has been selected as one of 4 prototype applications to reside on the new CEOS-Grid. Grid services such as the “Globus” toolkit provides security certificates and eventually HRM resources will be installed and prototyped on NCDC servers that will be linked to NASA, DOE, USGS, and the European Space Agency (ESA). The NESDIS Chief Information Officer (CIO) is supporting this effort at NCDC. For more information on NOMADS and CEOS-Grid. The NESDIS Chief Information Officer (CIO) funds this activity at NCDC. The activity has begun with the procurement of a server that will eventually move NOMADS data across the grid, using the Globus toolkit. For more information on Globus see: <http://www.globus.org>.

### **Funding Approaches**

What is needed to achieve significant progress toward making it easier to publish and use our distributed data assets under NOMADS? Currently, there are multiple independent efforts going on within NOAA and elsewhere that are testing different approaches and component pieces needed for a comprehensive set of data services. A first step in achieving the vision is to simply make involved researchers and developers aware of the other similar efforts within the institution.

Coordination and focus rather than more funding will be needed to assess where we are in achieving the vision of providing a rich set of distributed data services that can connect to each other to form a whole that is more than the sum of its parts. The NOMADS Team will do this by organizing regular seminars and meetings to present current efforts and plans and to discuss the important issues arising out of current efforts. The presentations and discussions must also include advances outside of the institution to provide scientific data frameworks and services, because NOAA's data assets must fit into such external contexts to be widely useful. A minimal funding resource to support critical NOMADS services such as OPeNDAP, should be investigated.

During the past two years, ESDIM pilot funds (NES\_02-444E), and locally provided resources at NOMADS core institutions, have provided on-going development and support for NOMADS. Specific funding documents are being generated independently for submission to the 2005 budget process, however program resource requirements for expanded capabilities are provided should resources be identified to advance NOMADS capabilities prior to this date.

Collaboration, coordination, and focus rather than significant funding is required to implement and sustain this already successful pilot program. The Director's of core NOMADS Centers (GFDL, NCDC, NCEP) have minimal base resources already in place, for the pilot program. For operations, additional resources are required.

### **Operational Staffing Requirements**

For operations, a three-tier approach is required: (Note: Tier 1b and c, are an expanded NOMADS capabilities if base resources become available).

<b>Activity</b>	<b>Operating Center</b>	<b>Resources</b>
Tier 1: Long-term Archive Services	NCDC	
a. Data and metadata management.		Three staff positions

- i. Hardware requirements
  - b. US National HelpDesk Six staff positions
  - c. Scientific Research and Project Management Two staff positions
- 2) Tier 2: Real-time Data Access Component Note 1. Two per Center
- 3) Tier 3: Collaborating Institutions, and International Programs Note 2. (as determined)

Note 1: Core NOAA NOMADS Participants: (CDC, GFDL, NCEP, PMEL)

Note 2: See Section XXXX for current participating institutions and programs.

### External Efforts

The NOMADS team is actively partnering with existing and development activities including the Comprehensive Large Array Stewardship System (CLASS); the National Oceanographic Partnership Program's (NOPP) National Virtual Ocean Data System (NVODS); the Department of Energy's Earth System Grid (ESG); and the Thematic Real-time Environmental Data Distributed Services (THREDDS) project being developed through the National Science Foundation and Unidata. Science and Technical Advisory Panels have been established and will ensure the NOMADS architecture can provide necessary inter-operability; and develop data archive requirement recommendations to NOAA. NCDC has taken the lead in this already successful collaborative effort with over 20 participating institutions (see Section XXX). This partnership for environmental prediction provides a much larger return on its total scientific investment dollar by leveraging core research being performed by Universities and other government laboratories, back into the operational NWP and climate model development and evaluation that is eventually provided back to the public for the protection of life and property, and for climate model assessment, impacts, and diagnostic purposes.

There exists a ground surge of distributed data access projects. The number and scope of these projects clearly demonstrate the viability of large-scale multi-discipline research in the physical sciences. Today, each data center or scientist tends to develop its own data sets and projects independently. While NOMADS places no formal constraints upon participants to achieve individual goals, it does provide direction, and suggested data formats and communications protocols to enable the sharing of interrelated data. The level of work required to provide for this interoperability depends upon the scope of the data provider's operations. However, a typical data center can implement and maintain these services with approximately 2 to 3 staff.

### NOMADS Participants and Data Suppliers

NOMADS data participants and suppliers currently include:

- NCDC
- PMEL
- NCEP
- GFDL

- CDC
- NCAR
- PCMDI
- USGODAE
- COLA (and GMU through COLA)
- NASA/GCMD (metadata and data locations)
- International (BADC, e-science, others in process)

### **Potential Data Suppliers**

Potential data suppliers include:

- Other data centers (NESDIS, NWS, NOS, NMFS, Unidata, etc.)
- CLASS
- IOOS
- GOOS
- GCOS
- CEOP
- CEOS
- Grids: NERC- Data Grid, e-science, IPG, ESG
- Other International:
  - 1) GODIVA (Keith Haines) Reading
  - 2) ESMF (modeling code framework)
  - 3) PRISM (NOMADS like activity in Europe)
  - 4) ESG (Dean Williams)
  - 5) BOM (Australia thru GODAE)
  - 6) BADC (Bryan Lawrence)
  - 7) CEOS-Grid (Rutledge, Diamond)
  - 8) Project CEOP (Japan)
  - 9) PyClimate (University of the Basque Country)
  - 10) Climateprediction.net

### **Application Programming Interfaces**

The core Application Programming Interfaces (API's) for NOMADS include:

- JAVA-OPeNDAP
- C++-OPeNDAP
- NetCDF
- THREDDS
- Python and Climate Data Analysis Tools (CDAT)
- Eventually include the work of the BADC for Grid based direct GRIB and BUFR interface.

### **Communications Protocols**

The core NOMADS communications protocols are:

- OPeNDAP
- HTTP

- FTP
- LDAP
- THREDDS
- XML
- Globus/CEOS-Grid/ESG

### **Data Conventions under NOMADS**

The following are a listing of data conventions currently available under the NOMADS framework:

- COARDS
- cf
- GRIB
- BUFR
- GRIB2
- FGDC
- HDF
- Non-COARDS compliant data sets can be accessed
- ascii (using ftp) services available.

NOMADS will embrace any convention that the OPeNDAP framework adopts. Some data forms we expect to use in the future include the OpenGIS convention.

### **Collaborators**

The current participating collaborators under the NOMADS framework include:

- NOAA National Climatic Data Center (Project Lead)
- National Weather Service, National Centers for Environmental Prediction (co-PI)
- NOAA Geophysical Fluid Dynamics Laboratory (co-PI)
- National Center for Atmospheric Research (co-PI)
- NOAA-CIRES Climate Diagnostics Center (co-PI)
- NOAA Forecast Systems Laboratory
- NOAA Pacific Marine Environmental Laboratory (co-PI)
- Center for Ocean-Land-Atmosphere Studies (co-PI)
- University Consortium for Atmospheric Research Unidata Program
- National Severe Storms Laboratory (collaborating with the University of Wisconsin, Space Science Engineer Center)
- NASA Global Change Master Directory
- NASA Seasonal-to-Interannual Earth Sciences Information Partner
- LLNL Program for Climate Model Diagnosis and Intercomparison (co-PI)
- DOE Earth System Grid

- George Mason University
- University of Alabama, Huntsville
- University of Washington
- University of Iowa
- Committee for Earth Observing Satellites (CEOS) Grid Project

## **DATA Availability**

The distributed NOMADS framework will provide data manipulation capabilities and real-time and retrospective access to NWP model input and output, ensembles, reanalysis, GCM's and observational data both upper-air and surface based, including oceanographic measurements being served by OPeNDAP enabled hosts. NCEP model output will be ingested at NCDC through NCDC's NOAAPort Data Access and Retrieval System (NDARS, Rutledge, et al., 2000) in real-time. These data have been archived at NCDC since October 1999. Real-time NWP data will also be available through NCEP's real-time NOMADS data services effort (Alpert, et al., 2002).

NCEP model output grids that will be available in 2003 include the Global Forecast Model (GFM, formally called AVN and MRF), the Rapid Update Cycle (RUC), Eta, mesoEta, NGM and the Weather Research Forecast (WRF) model when available. Also available under NOMADS will be NCEP's model input data and "run history" information including model spectral coefficients. GFDL, NCAR, and other climate simulations and diagnostics will also be available within the framework. For currently available NWP and other data see the newly established NCDC site at <http://nomads.ncdc.noaa.gov>. For GCM data currently available see the GFDL NOMADS site at <http://nomads.gfdl.noaa.gov>. Near real-time GDAS model input data will be available for testing and limited user access by the spring of 2003.

At NCDC archive and volume requirements will hinge on

- 1) the volume of the NOAAPort AWIPS grids (currently approximately 1TB/yr);
- 2) the volume of the full suite of NCEP Global and meso grids (approximately 100TB/yr)
- 3) (Note: Full resolution grids will be archived no longer than 5 years)
- 4) the volume of the NCEP GDAS model input data set (approximately XXX TB/yr) and
- 5) the volume of expected for long-term archive of GFDL's climate models (approximately 2TB)

## **Real-Time NOMADS NWP Services at NCEP**

The NCEP Real-time NOMADS (RT-NOMADS) prototype project serves real time operational data only. NCDC is the operational archive focal point and project leader for NOMADS and holds data sets older than real time. The NCEP RT-NOMADS server sends NCDC data for archival. The NCDC NOMADS goal is to save model (run history) data but budget realities provide that the archives can save initial conditions and observations sufficient to restart NCEP models to reconstruct model run history. A NOMADS goal at NCDC is to add to this archive, the NCEP operational run history forecasts as soon as sufficient storage is funded. Other NOMADS participants serve their own data sets. The NOMADS participant data sets will be

connected by various search engines now under development at UCAR (THREDDS), NCAR, and NASA. The RT-NOMADS project is a prototype with the goal to produce an Operational Specification which if built is planned to be located in the operational data distribution component of NWS.

### **NCEP Model Input Data Availability**

The NCEP Global Data Assimilation System (GDAS) analysis files will be ingested through NCDC Load Balanced System and will be available under NOMADS from the NCDC archive. Data are currently documented in Federal Geographic Data Committee (FGDC) format as NOAA required. This documentation can be accessed at NCDC under Tape Deck (TD) No. 6172 at: [www4.ncdc.noaa.gov/ol/documentlibrary/datasets.html](http://www4.ncdc.noaa.gov/ol/documentlibrary/datasets.html).

The GDAS dataset consists of the minimum set necessary to re-generate NCEP analysis and forecast products (model re-start and initialization files). GDAS includes the Global Spectral Forecast Model (GSM), and the Spectral Statistical Interpolation (SSI) Cycling Analysis System (CAS) with triangular truncation (T) 170 and 28 sigma levels. To start the CAS, model spectral coefficients are provided on gaussian grid in a sigma vertical coordinate system. These data represent the model's "ground truth", and the best estimate- in terms of analyzed fields- for scientific study. Data that are restricted may not be available.

"Post" is a FORTRAN program is available from NCEP that will convert spectral coefficients to gaussian grid, sigma to pressure, and gaussian to latitude and longitude. Work continues to couple Post to the NOMADS user interface for source and executable downloads. The GDAS dataset under NOMADS will include the Global Spectral Forecast Model (GSM) and the Spectral Statistical Interpolation Cycling Analysis System (SSI-CAS) binary files and contains ~2.5Gb per day (4 cycles/day: 00Z, 06Z, 12Z, and 18Z). The binary files are raw data, which are acted on by NOMADS servers to produce useful grids. The archived analysis data sets serve as model verification as well as the source for model reruns and retrospectives. Including the observations allows for cycling analysis systems to re-analyze the observations. Never before has this model input data and information been available to the public. A partial list of NOMADS planned available observations (with associated data format) include:

- Analysis Bias Corrected Information (ASCII)
- Ship / Buoy Observations (BUFR)
- Guess prep / guess output (BUFR)
- Observational Toss List (ASCII)
- Bogus Observations (BUFR)
- ACARS and Aircraft (BUFR)
- Wind Observations (BUFR)
- Analysis Ready Obs. (prepBUFR)
- Surface Analysis Restart Files (BUFR)
- Surface and Upper-Air observations (BUFR)
- Fixed Snow Field (GRIB)

Previous 6 hour forecast (BUFR)  
“Post” Guess Output (spectral binary)  
Profiler (BUFR) / SST’s (GRIB)  
MSU 14 and HIRS 14/15/16 (IEEE)  
SSM/I Satellite obs (BUFR)  
NOAA (satellite) 15/16 AMSU - A/B  
TOVS 1B Radiances (IEEE)  
TOVS Satellite Obs (BUFR) / GOES Satellite Obs (BUFR)  
O3 Sat Obs (binary) and ERS Sat obs  
SBUV: Satellite Wind Observations  
Radar VAD Winds (BUFR)

The formats of these data sets are generally dictated by the necessity to run models efficiently on modern computers. NOMADS converts the formats and structure to the users requested form but also allows the raw data to be directly accessed.

### **Introduction to Four-Dimensional Data Assimilation Approach (4DDA)**

In order to develop international systematic approaches to model evaluation, and for integrating various earth system models, (coupled atmospheric-ocean, sea ice, carbon cycle, etc.), an expanded global monitoring and climate change detection must address the issues of data homogeneity, heterogeneous measurement systems, calibration differences across and among sensor platforms and other inconsistencies that inhibit a dynamically consistent time series.

Over the last decade a unified approach for combining and synthesizing data has been adopted in the form of the Four-Dimensional Data Assimilation (4DDA) approach by numerical weather prediction centers namely NCEP and ECMWF. The 4DDA approach provides temporally and physically consistent global analysis. It incorporates differing input data in a spatially and temporally consistent manner and interpolates these data to fill in gaps in the background field, and even for missing variables. This approach, especially the use of reanalysis data sets where satellite data plays an increasing role in assimilation simulations can be rerun many times is valuable for quality control. A major limitation of this approach is that the analysis can be quite sensitive to the quality of the model used. A major benefit is that this is a consistent analysis that can be run and re-run with varying input parameters or longer term forcing (Mahlman, 1995).

The data available under NOMADS include the NCEP and NCAR "Reanalysis". This global analysis is a 40-year record of global analyses of atmospheric fields in support of the needs of the research and climate monitoring communities. This effort involves the recovery of land surface, ship, rawinsonde, pibal, aircraft, satellite and other data, quality controlling and assimilating these data with a data assimilation system, which is kept unchanged over the reanalysis period 1957 through 1996. This eliminates perceived climate jumps associated with changes in the data assimilation system.

### **NOMADS Model Rerun and Retrospective Capability**

To accomplish a retrospective and rerun capability of NCEP models NOMADS intends to save what is necessary to rerun a forecast model to regenerate products as close as possible to NCEP Operations. That includes all observations, fixed fields, and Initial condition or restart files. This means that an analysis cycling system can be run or a forecast model can be run from initial conditions that are saved on the NOMADS archive.

Even as storage costs have decreased the volume of model data, which is central to research and development has increased at a more rapid pace. Programming resources are needed to organize software to rerun the models to automate data set creation. The run history data will be retrieved and available by utilizing CPU resources at appropriate model test bed facilities to create requested data sets on demand for scientific study and model development. Program code to rerun NCEP models will be supplied for interested users to make their own reruns. Many Completed retrospective reruns are archived in response to community interest.

To start the operational Spectral Statistical Interpolation (SSI) a number of standard observation files, such as RAOB's are necessary as well as less conventional observations such as satellite, radar, ACARS, etc data. The restart files, which in this case are equivalent to analysis, as well as initial condition files over the globe contain spectral coefficients on gaussian grid in the sigma vertical coordinate system. These files represent a ground truth, or the best estimate in terms of analyzed fields for scientific study. Therefore, a *post* program to convert spectral coefficients to gaussian grid, *sigma to P* (pressure) conversion program, and *gaussian to longitude/latitude* conversion is included in the package and coupled to the NOMADS GUI as well as executable and source code for a number of platforms. Work continues on including these utilities in the server systems as well as the web based ftp services.

### Direct *ftp* Retrieval of Data Sets

Direct *ftp* Retrieval of Data Sets is used for large blocks of data. LAS and DODS/OPENDAP are based on accessing a high level of granularity, which can become inefficient for large data set transfer. The NCEP service [ftp2u](#) (see Section XXXX) provides for large array downloading.

### NCEP Model Data Volume

The latest 1-2 years of NCEP minimum data set for model retrospective and rerun of the GDAS will remain on servers at NCEP. An ongoing archive procedure will deliver NCEP data sets to the NCEP NOMADS server located at NCDC. GDAS fields will be permanently archived at NCDC.

Volume assignments for the NCEP model minimum set for the rerun and retrospective archive are shown in the table below:

<u>Model</u>	<u>Phase I</u> Year 1	<u>Phase II</u> Year 2	<u>Phase III</u> Year 3
Global	500GB/year	1TB/year	10TB/year

Meso	800GB/year	3TB/year	30TB/year
Ruc/Ocean...	500GB/year	1TB/year	30TB/year
<b>TOTAL</b>	<b>1.8TB/year</b>	<b>5TB/year</b>	<b>70TB/year</b>

## **The Weather Research and Forecast Model**

NOMADS is a technology for collaboration and model development as exemplified by the Weather Research and Forecast (WRF) model. The contribution of the scientific community to the WRF development effort requires the transfer of data sets of model retrospective results for testing and comparison. NOMADS enables both retrospective and real time access to the suite of digital products from reanalysis and operational results to give a range of users, from commercial to university scientist, model comparison access and supply operational grade initialization for forecast model study. For more information regarding the emerging WRF model see [www.wrf-model.org/documentation\\_main.html](http://www.wrf-model.org/documentation_main.html)

## **NOAAPort AWIPS Grids at NCDC**

As described in Section XXX, NCDC will ingest and archive the entire suite of NWP gridded output fields delivered over NOAAPort. These data will be available on-line for at least 6 months, and then cycled to the NCDC archive for long-term storage (not to exceed 5 years). Table XXX summarizes the current NOAAPort data stream of NWP products. Specific holdings including model parameters (e.g., temp, ht, u/v ,etc.) can be obtained at the NCDC NOMADS Web Page at <http://nomads.ncdc.noaa.gov>.

Table XXX. TBD

## **NCEP Regional Reanalysis**

By early 2003 (March), the new Regional Reanalysis will also be available under NOMADS at NCDC and other NOMADS sites. Further information and data access links will be provided when the data become available.

Under NOMADS and in GRIB format, lambert-conformal grids of the new Regional Reanalysis (analyses only) with a temporal resolution of 4x daily will be available with a volume of approximately 3 TB for the period of record (1978 to the present). It is expected these files will be updated monthly. The Regional Reanalysis is expected to become available from NCEP in the Spring of 2003.

NCEP is developing software to make regional subsets of lambert-conformal GRIB files. When completed, users will be able select a subset of the ETA 12 km forecast for downloading. This software will be compatible with the Regional Reanalysis files. NCEP along with COLA, will develop the GrADS capability of NOMADS to use these data.

## **GFDL GCM Data Availability**

General Circulation Models available under the NOMADS framework include the GFDL R-30 climate model. R-30 is a coupled Atmosphere-Ocean General Circulation Model (AOGCM). Its four major components are an atmospheric spectral GCM, and ocean GCM, and relatively simple models of sea ice and land surface processes. The name "R30" is derived from the resolution of the atmospheric spectral model (rhomboidal truncation at wave number 30). The R30 model is identified as GFDL\_R30c in the Third Assessment Report (TAR) of the Intergovernmental Panel on Climate Change (IPCC). See Chapter 9, Table 9.1 of "Climate Change 2001: The Scientific Basis" IPCC Working Group I (2001).

The model output that is stored on the GFDL NOMADS server are taken from six experiments conducted at GFDL using the GFDL\_R30c model. Data files produced by one long-running control integration (one with no changes in external forcings, e.g., constant CO<sub>2</sub>) and five climate change scenario experiments have been made accessible to interested researchers. More information regarding the individual experiments is found in the Control & Transient Forcing Experiments section of the GFDL NOMADS Web site. The GFDL R-30 numerical model was developed and the experiments conducted by members of the Climate Dynamics and Prediction Group at the GFDL in Princeton, New Jersey. For further information on current available GCM data see the NOMADS site at GFDL at <http://nomads.gfdl.noaa.gov>.

The following is an evaluation of what climate data (greenhouse warming focused) GFDL will contribute initially to NOMADS.

### **1. Products**

The data from climate model studies of warming due to increases in the greenhouse gases is the main focus here. The model used is a coupled model consisting of an atmosphere, ocean, land surface and sea ice components. Coupled models are used to study historical climate variations on decadal and longer time scales. These models are also forced with estimates of the future radiative forcings to study future climate changes.

It should be noted that this type of experiment is mainly a boundary value problems, while NWP is mainly an initial value problem. The experimental design is normally a control integration is preformed where there are no long term changes in the radiative forcing. Perturbation experiments are then done off that control integration. The control integration allows studies of the natural or unforced internally generated variability of the climate system. Comparing the perturbation results with the control allows one to assess the model response to the changes in the forcing.

A table of the currently available model data sets follow:

GFDL\_R15\_b - Low resolution (R15L9 atmosphere, 4 degL12 ocean)

A. Control integration 1000 to 1500 years

B. 9 IS92A integrations (ensemble) - 1766-2095; 330 years each

C. 1 1% CO<sub>2</sub> increase - 80 years

D. 1 0.5% CO<sub>2</sub> increase - 150 years

GFDL\_R30\_c - medium resolution (R30L14 atmosphere, 2 degL18 ocean).

E. Control integration 900 years

F. 3 IS92a (ensemble) 1866-2095; 230 years each

G. 1 SRES A2 1990-2100; 110 years

H. 1 SRES B2 1990-2100; 110 years

I. 1 1% CO<sub>2</sub> increase, 80 years

Notes:

IS92a - a IPCC radiative forcing scenario from 1990 to 2100. Estimates of historical man-made forcing were used 1766 to 1990, "natural" forcings (solar and volcanic) were not used in this integrations. SRES A2 and B2 are newer IPCC estimates of the future radiative forcings  
1% CO<sub>2</sub> increase integrations are very useful for model intercomparisons

## 2. Services to provide/users

There are a number of assessment for future climate changes and their impacts going on the around the world. By far the most important of these is the on going IPCC (Intergovernmental Panel on Climate Change). The IPCC is currently in its third assessment of climate change science. Recently, the US just completed its first National Assessment. GFDL has contributed data to both of these assessments (in addition to many others).

The various model intercomparison projects comprise a second group of users, many under WMO oversight. The intercomparisons seek to understand the causes for differences in the model's response to various forcings and to document the successes and failures of the models. For climate change research, the Coupled Model Intercomparison Project (CMIP) is the most comprehensive.

A third group of users are the various scientists who perform analysis on the publicly available model data. Currently, NCDC provides GFDL model output to the public. A number of important scientific discoveries have resulted from collaborations with this group of users.

Finally, commercial users are interested in this type model model data, particularly those which have long term investments in infrastructures that are climate sensitive. Some examples of this type of user include water resource management (dam planning), and agricultural interests.

## 3. Data Availability (using the table from part 1 above):

- A. 39 GB for monthly data (1000 yrs)  
120 GB for daily data (100 yrs)
- B. 13 GB for monthly data (330 yrs)  
48 GB for daily data (40 yrs)
- C. 4 GB for monthly data (100 yrs)  
24 GB for daily data (20 yrs)
- D. 6 GB for monthly data (150 yrs)

- 24 GB for daily data (20 yrs)
- E. 240 GB for monthly data (1000 yrs)  
400 GB for daily data (100 yrs, 4 20 yr periods)
- F. 55 GB for monthly data (230 yrs, per ensemble member)  
200 GB for daily data (2 20 yr periods, per ensemble member)
- G. 20 GB for monthly data (110yrs)  
100 GB for daily data (1 20 yr period)
- H. 20 GB for monthly data (110yrs)  
100 GB for daily data (1 20 yr period)
- I. 20 GB for monthly data (100yrs)  
100 GB for daily data (1 20 yr period)

Total - Approximately 2TB

### **NCAR Data Availability**

NCAR also has a distributed computing effort collaborating under NOMADS. Working with UCAR and NCAR, NOMADS will partner with the forward-looking pilot project called the Community Data Portal (CDP). The CDP is targeted directly at elevating NCAR's collective ability to function as a data provider with a coherent web-based presence. Under the CDP it is expected that portions of Community Climate System Model (CCSM), and Parallel Climate Model (PCM), NCEP reanalysis and other data will be available under the NOMADS framework. A NOMADS GDS server has been established at NCAR and can be reached at <http://motherlode.ucar.edu:9090/DODS/OPeNDAP/>.

### **Climate Diagnostic Center Data**

NOAA's Climate Diagnostic Center (CDC) a NOMADS collaborator with an extensive array of distributed data sets for Web access. CDC is currently has one of the largest inventories of client-server listing in the world. CDC is well known for "one-stop shopping" for NCEP reanalysis, and many other data sets and observations. For more information on the CDC data sets see [www.cdc.noaa.gov/PublicData/data\\_descriptions.html](http://www.cdc.noaa.gov/PublicData/data_descriptions.html).

### **COLA**

The Center for Ocean-Land-Atmosphere (COLA) has an extensive distributed climate and weather grids available under the GrADS Data Server. See <http://cola8.iges.org:9090/index.html> for a listing of available data under the NOAMDS framework.

### **NASA's Global Change Master Directory**

NASA's Global Change Master Directory (GCMD) is a NOMADS collaborator and provides descriptions of Earth science data sets and services relevant to global change research. The

GCMD database includes descriptions of data sets covering agriculture, the atmosphere, biosphere, hydrosphere and oceans, snow and ice, geology and geophysics, paleoclimatology, and human dimensions of global change. The DODS/OPENDAP portal at the GCMD can be reached at

<http://gcmd.gsfc.nasa.gov/Data/portals/DODS/OPeNDAP/index.html>.

### **Data Availability at Lawrence Livermore National Laboratory**

Data sets that are currently available under the NOMADS framework at LLNL include:

AMIP I

AMIP II

CMIP I

CMIP II

NCEP/NCAR Reanalysis

NCEP/DOE Reanalysis

Data sets currently planned include:

ERA-15

PCM

CSM

Climate change detection archive (similar to MPI's)

CMIP II+

ERA-40

### **Other data centers data sets underdevelopment (with review of this document)**

### **The Future of NOMADS**

The long-term viability of NOMADS will be directed toward the collaborative nature of the current NOMADS partners, and future expected projects particularly the Earth System Grid, Ceos-Grid, and the British Atmospheric Data Center (BADC). BADC will provide limited access to the European community suite of data including ECMWF, Hadley and eventually UKMET. Australia and Japan are already participating under USGODAE, and CEOP (not yet a firm CEOP commitment). A vision currently taking shape is that an umbrella of services available to scientists. This umbrella will include NOMADS services, NCAR data portal, the Earth System Grid (ESG), the CEOS-Grid, NERC DataGrid (UK), and other efforts. Under an open source Python-based operating environment, NOMADS users will be able to use whatever tool they so desire and the NOMADS-ESG will allow users to intercompare data of differing formats and from differing locations using either thin-client (Web browser), or up to thick-clients (host services using CDAT, and/or Internet2). Once we have this framework, day-to-day maintenance will be secured 1<sup>st</sup> via NOAA level initiatives then a cross line office activity for minimal resources will be initiated by the NOMADS Steering Group.

Finally, an “open source” modular design is being considered for integration into the NOMADS core infrastructure. This will initially include the CDAT tools that are Python based. In this way, NOMADS users can have access to discipline independent libraries of user contributed modules of equations and formulas, statistical routines, and other tools as developed in other

scientific disciplines such as Physics, Mathematics, or Medicine. Once such project PyClimate, already exists and is under the NOMADS umbrella through a LLNL collaboration with the PyClimate developers at the Department of Applied Physics II, Faculty of Sciences, University of the Basque Country (<http://starship.python.net/crew/jsaenz/pyclimate/>).

## **Future Requirements**

Even with the emergence of new web based services, the comparison of GCM results with the observational climate record is still difficult for several reasons. One limitation is the global distributions of a number of basic climate quantities, such as precipitation or clouds, are not well known. Similarly, observational limitations exist with model re-analysis data. Both the NCEP/NCAR (Kistler, Collins, Kalnay, et al., 2001) and the ECMWF (Gibson, et al., 1997) re-analysis eliminate the problems of changing model analysis systems but observational data also contain time-dependant biases by changing observational networks, station moves, and the assimilation of various remotely sensed data (Rutledge, et al., 1991) using differing sensor instruments, or calibrations. These changes in input data are blended with the natural variability making estimates of true variability uncertain. The need for data homogeneity is critical to study questions related to the ability to evaluate simulation of past climate. One approach to correct for time-dependant biases and data sparse regions is the development and use of high quality “reference” data sets (Karl, et al., 2000).

Beyond the ingest and access capability being implemented with NOMADS are the challenges of algorithm development for the inter-comparison of large-array data (e.g., satellite and radar) with surface, upper-air, and sub-surface ocean observational data. The implementation of NOMADS will foster the development of new quality control processes by taking advantage of distributed data access. One major challenge facing the scientific community is the development of methodologies for the inter-comparison of large-array observational data sets with model simulations. In the near future, NOMADS would include the development of algorithms to blend remotely sensed data with in-situ surface and upper-air data, then to use these blended fields for verification and validation of both weather prediction and climate models in both time and space (Rutledge, et al., 2001).

The interface between the NOMADS spinning disks, and hierarchical tape archive systems such as NCDC’s mass storage systems, will be developed during 2003/2004 so users may obtain access to data sets destined for long-term storage.

## **U.S. National Model “HelpDesk”**

Under President Bush’s new Climate Change Research Initiative (CCRI), access to and understanding of climate models and data is a high priority. Several Agencies including the U.S. Departments of Agriculture, Commerce, Energy, and the EPA and NOAA, are currently collaborating to monitor emissions, foster international partnerships for research, uncertainty reduction in climate models, and develop technologies for access to climate models for large-scale systematic approaches to model evaluation. A large and growing number of impacts groups and carbon cycle researchers will require access to, and understanding of these models

and data. NCDC is charged with the archive of these data, however users at all skill levels will require on-line assistance with easy to navigate Web interfaces to model data. There is a danger that if users are not properly educated on the science or use of a specific model, it may be interpreted incorrectly, or users may use a model in a way that is inappropriate. Users must not only have access to data but the science and information behind the models, and how to properly use these complex data sets.

In support of these activities and the NOMADS project itself, NCDC has proposed the formation of a “U.S. National Climate and Weather Model *HelpDesk*” to assist these model users. The *Helpdesk* will:

- Provide a tailored Web interface to NCDC’s archives of climate and weather models and
- a capability to service users with the information they need, not just model output and data.
- Develop distributed database search engines.
- Provide for the long-term stewardship of these high volume and complex data and metadata at NCDC and elsewhere in a distributed framework.

## **Bibliography**

Alpert, J.C., 2002, Rutledge, G.K., Williams, D., Stouffer, R., Buja, L., Doty, B, Hankin, S., Domenico, B., Kafatos, M., 2002, “The Plan to Access Real-Time NWP Operational Model Data Sets using NOMADS”, Proceedings of the 18th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, American Meteorological Society, Orlando FL. 73 pp.

Bray, T., Paoli, J., Sperberg-McQueen, C.M., 1998, “Extensible Markup Language (XML) 1.0 Specification”, W3C REC-xml-19980210: [www.w3.org/TR/1998/REC-xml-19980210](http://www.w3.org/TR/1998/REC-xml-19980210).

Davis, E. R., J. Gallagher, J., 1999, “Using DODS/OPENDAP to Access and Deliver Remote Data”, Proceedings of the 15th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, American Meteorological Society, Dallas, TX. 571 pp.

Doty, B.E., Wielgosz, J., Gallagher, J., Holloway, D., 2001, “GrADS and DODS/OPENDAP”, Proceedings of the 17th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, American Meteorological Society Albuquerque, NM. 385 pp.

Foster, I., Kesselman, C. and Tuecke, S. 2001, “The Anatomy of the Grid: Enabling Scalable Virtual Organizations”, International Journal of High Performance Computing Applications, 15 (3). 200-222.

Gibson, J. K., P. Kallberg, S. Uppala, A. Noumura, A. Hernandez, and E. Serrano, 1997, “ERA

Description. ECMWF Re-Analysis Project Report”, Series 1, ECMWF, Reading, UK, 77 pp.

Hankin, S., J. Callahan, and J. Sirott, 2001, “The Live Access Server and DODS/OPENDAP: Web visualization and data fusion for distributed holdings”, 17th Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, American Meteorological Society, Albuquerque, NM. 380 pp.

Karl, T., D. Easterling, P. Groisman, et al., March 2000, “Observed Variability and Trends in Extreme Climate Events: A Brief Review”, Bulletin of the American Meteorological Society, 81, 417 pp.

Kistler, R., W. Collins, E. Kalnay, R. Reynolds, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, J. Janowiak, M. Kanamitsu, K. Mo, C. Ropelewski, R. Jenne, D. Joseph and M. Fiorino, 2001: The NCEP/NCAR 50-year Reanalysis: Monthly-means CD-ROM and Documentation. Bulletin of the American Meteorological Society, 92, 247 pp.

Mahlman, J.D., 1995, “Toward a Scientific Centered Climate Monitoring System”, Climatic Change 31: 223-230.

NCAR Strategic Plan for High Performance Simulation, May 19, 2000. Not published.

Ramachandran R., M. Alshayeb, B. Beaumont, H. Conover, S. Graves, X. Li, S. Movva, A. McDowell and M. Smith, 2001: "Earth Science Markup Language: A Solution for Generic Access to Heterogeneous Data Sets," Earth Science Technology Conference, Maryland

Rutledge, G.K., E. Legg, and P. Menzel, 1991, “Operational Production of Winds from Cloud Motions”, Palaeogeography, Palaeoclimatology, Palaeoecology, Vol. 90 No. 1-3 (Global and Planetary Change Section), 141 pp.

Rutledge, G.K., A. Stanley, E. Page, L. Spayd, and J. Brundage, 2000, “Implementation of the NOAA Port Data Archive and Retrieval System (NDARS) at the National Climatic Data Center”, Proceedings 16<sup>th</sup> Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Long Beach, American Meteorological Society, 492 pp.

Rutledge, G.K., T. Karl, D. Easterling, L. Buja, R. Stouffer, 2001 “Evaluating Transient Global and Regional Model Simulations: Bridging the Model/Observations Information Gap”, (Invited), American Geophysical Union, Spring Meeting Boston MA S35 pp.

Rutledge, G.K., D. Williams, R. Stouffer, J. Alpert, L. Buja, B. Doty, S. Hankin, B. Domenico, M. Kafatos, 2002, “The NOAA Operational Model Archive and Distribution System (NOMADS)”, Proceedings 13th Symposium on Global Change and Climate Variations, American Meteorological Society, Orlando FL. J76 pp.

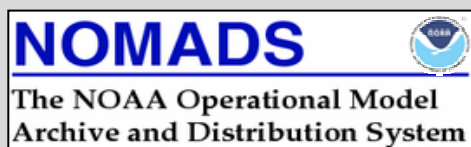
Trenberth, K.E.: 1995. “Atmospheric Circulation Climate Changes”, Long-Term Climate

Monitoring by the Global Climate Observing System 297-323; and Climatic Change, 31: 427-453

Williams, D.N., R.S. Drach, P.F. Dubois, C. Doutriaux, C.J. O'Connor, K.M. AchutaRao, and M. Fiorino, 2002: Climate Data Analysis Tool: An open software system approach. 13th Symposium on Global Change and Climate Variations, American Meteorological Society extended abstract volume, Orlando, Florida, J71pp.

Yang, R., X. Deng, M. Kafatos, C. Wang and X. Wang, 2001, "An XML-Based Distributed Metadata Server (DIMES) Supporting Earth Science Metadata", Proceedings of the 13th International Conference on Scientific and Statistical Database Management, Institute of Electrical and Electronics Engineers, Computer Society, Fairfax VA. 251 pp.

**The  
NOAA  
Operational Model Archive and Distribution System (NOMADS)  
Data Management Vision**



**A Data Management Plan for use by NOMADS Participants and  
by NOAA IT Managers**

**DRAFT**

**NOAA Core Collaborators**

Climate Diagnostic Center  
Forecast System Laboratory  
Geophysical Fluid Dynamics Laboratory  
National Climatic Data Center  
National Centers for Environmental Prediction  
Pacific Marine Environmental Laboratory

**Core External Collaborators**

COLA  
DOE/LLNL/PCMDI  
NASA/GCMD  
UCAR/NCAR  
UCAR/Unidata  
BADC  
BOM

**Glenn K. Rutledge, Principle Investigator**

**January 12, 2003**

## Part II.

### Data Management Vision under the NOMADS Philosophy

#### Data Management Vision

Making our valuable data collections visible, widely available and linking them to other, external, data sources adds a new method for advancing scientific discovery and data exploration. In addition to theory, experiment, and simulation, data exploration only becomes possible when it is easy to browse, access, and combine data from multiple sources within a single application.

From standard data analysis and comparison tasks, to data assimilation, fusion and mining, and into digital libraries and beyond, new online data applications are evolving at a very rapid rate. NOAA's integrated data management plan is occurring in a context of significant efforts by other organizations to provide portals, catalogs, and gateways to environmental data resources (National Environmental Data Index (NEDI), Global Change Master Directory, Gateway to Global Change Data, Master Environmental Library (MEL), Earth Observing System Data Gateway, GeoConnections, FGDC Clearinghouse, USGS "Gateway to the Earth", etc.). It is clear that it is no longer sufficient for any one national center to develop its data services alone. Both researchers and policy-makers alike now expect our national data assets to be easily assessable and interoperable with each other, regardless of their physical location. The effective interagency distributed data services implied by this require coordination of data infrastructure and management extending beyond the organizational boundaries of any individual center.

NOMADS will provide access to a number of basic environmental monitoring data sets. A distributed framework such as NOMADS installs the necessary access methodology for access to these data, however this document, or the NOMADS program itself does not intend to track and manage all the data sets across the "networks". Several programs are currently being developed for a dynamic engine to data cataloging, and for user search and retrieval. Once such program is called THREDDS. A multi-instituted program, THREDDS will:-----

This document will outline the core datasets currently available in terms of broad categories, and intended uses for selected and new data sets. However, any attempt to classify and document and maintain the varied data sets under NOMADS would quickly become out of date and non-functional to the intended goals of NOMADS. Each data provider will document and maintain their own holdings, and maintain or revise as new datasets evolve.

The NOAMADS data management vision is to make data and information services uniformly available to all elements of the NOAA community with transparent wide area access to these services - and thus provide seamless provision of and access to actual data independent of how and where they are collected and stored. Services at this level will involve development of skill-appropriate interfaces (beginner or expert user, student, scientist or policy maker), security/authentication services to ensure quality assurance and data providence and sufficient data abstractions to enable wide-ranging data fusion and visualization.

In the not-too-distant future, access to data from analysis and visualization applications could be as simple as access to documents is today through a web browser. Just as our networks and web browsers now interoperate to support location transparency (not knowing or caring whether a document you access is stored locally or remotely), so our programs should be able to access, analyze, visualize, and integrate data from remote sources as if it were local. Just as HTML is now the lingua franca of the web that enables almost anyone to publish a document widely, so an XML-based metadata standard could provide the wrapper by which datasets could be published for access by applications and searched by data search engines.

Just as web browsers support multiple protocols for accessing text and images using various protocols (FTP, HTTP, ...) and formats (HTML, PDF, text, GIF, PNG, ...), so a remote data access infrastructure needs to be flexible enough to support multiple data access protocols (OPENDAP, ADDE, LAS, FTP, ...) and formats (netCDF, GRIB, HDF, ...). Just as legacy documents can be made available without converting them to a new format by dynamic generation of HTML from programs accessing a database, so legacy data can be made available to applications by dynamic generation of metadata wrappers and servers that can generate data subsets and metadata on-the-fly. Just as useful results are available from search engines searching plain text today without having a complete solution to the metadata discovery problem, so useful search and data discovery is practical without first finding a perfect representation for scientific metadata.

And just as the existence of the Web and browsers attracted people to begin using HTML and running servers, so a similar data access framework made from a few existing component protocols and formats that have already proved useful for remote data access could be put together into a framework that would attract users.

## External Collaborations

A number of data management efforts are currently being developed in collaboration with external organizations:

NCAR (ATD) with University of Oklahoma, University of Wisconsin, University of Washington, University of Hawaii, University of Minnesota, University of Illinois, University of Colorado, University of Wyoming, University of Denver, Penn State University, Colorado State University, NCAR with GFDL/PMEL/NCEP/NCDC/CDC via the NOMADS project  
NCAR (SCD) with DOE (The Grid) and numerous UCAR affiliated universities  
Unidata is collaborating with 83 U.S. universities, 18 .gov sites, 8 .coms, 8 international sites, 6 .mils (Internet Data Distribution system)  
SSEC, University of Wisc. (VisAD data model, multidimensional vis)  
National SMETE Digital Library (digital libraries, metadata)  
University of Rhode Island and MIT (DODS/OPeNDAP)  
SSEC (McIDAS analysis and visualization, ADDE remote data access server)  
NCEP (nawips, GEMPAK analysis and visualization); (NOMADS-RT)

Lamont Doherty Earth Observatory (THREDDS data servers)  
Australian Bureau of Meteorology (Java meteorological applications)  
University of Oklahoma, CAPPS (CRAFT project for distributing level 2 NEXRAD data)  
FNMOC (distributing NOGAPS model outputs to universities)  
NCDC (archiving NEXRAD level 2 data using LDMs)  
NOAA FSL (distribution of profiler and ACARS data)  
Meteorology Canada (distributing GEM model data)

## A Vision for Data Usability

Today, the primitive state of access to most NOAA data from real-time observations, field projects, model outputs, and long-term archives is similar in many ways to the state of access to documents before the advent of the Web, the world's largest and most successful distributed system. In pre-Web days, access to documents involved locating the documents, transferring the needed files, converting from one document format to another, and transforming document excerpts into common forms suitable for reading or merging into papers, proposals, presentations, or reports. Today, accessing data needed for research involves finding out where the data are stored, determining what specific files are needed, discovering which of many diverse formats are used, transferring the data to a local site, converting the data formats into whatever is required by local visualization and analysis applications, and possibly regridding the data for combination with other data.

In the not-too-distant future, access to data for analysis and visualization could be almost as simple as access to documents is today through a browser interface. Just as our networks and web browsers now interoperate to support location transparency, so our programs should be able to access, analyze, visualize, and integrate data from either local or remote sources. In the same manner that HTML has become the lingua franca of the web that enables anyone to publish documents, a standard metadata architecture could provide the means by which our important datasets could be easily published for access by local and remote applications, catalogued by search-engine services, and found by web browsers and other applications.

Web browsers now support accessing text and images using multiple protocols (FTP, HTTP, NNTP, SMTP, ...) and formats (HTML, PDF, text, GIF, RealAudio, etc.). Similarly, a remote data access infrastructure needs to be flexible enough to support multiple data access protocols (FTP, OPENDAP, LAS, SQL, ADDE, ...) and data formats (netCDF, HDF, GRIB, BUFR, OGIS, Excel, etc.). Just as legacy documents can be made available without converting them to a new format by dynamic generation of HTML from programs accessing a database, so legacy data can be made available to applications by servers that convert data slices on-the-fly and by dynamic generation of derived data in a form requested by a client application. Finally, given that it is practical today to find documents using search engines without a complete solution to the metadata discovery problem, so useful search and data discovery may be practical without first finding a perfect representation for scientific metadata.

The existence of the Web and browsers provided benefits that encouraged the use of HTML and web servers which multiplied as use of the web became more widespread. Similarly a "data

Web" constructed from existing component protocols and formats for remote data and metadata access could be combined into a framework that would increase benefits as usage increased. The data web will only become a reality when it is as easy to publish data with metadata and data services that make it as useful as it is to publish documents on the Web. NOAA might even play a role analogous to CERN in catalyzing such a vision by building the prototype infrastructure that demonstrates the practicality of the vision.

## **Achieving the Vision with Existing Resources**

It is not enough just to share the vision and make each other aware of what we are doing. Where reuse makes sense, we must find a way to take advantage of the efficiencies available in adapting general solutions to specific problems. This is difficult, because projects are not funded to provide general solutions that are reusable in other contexts, and the construction of generally useful frameworks is very challenging. Similarly, data providers usually have no incentive to provide the extra metadata and organization to their data that would make it useful in unanticipated contexts. How can resources be applied to leverage work in one area to make it available for use in other closely-related areas? As a specific example, how can resources be applied to further the development of a parallel version of widely-used data access software for use by modelers on supercomputers, when the institutional unit that maintains and develops the software is not funded to develop modeling software on supercomputers?

## **Strategy**

### **Overview**

NOAA's integrated data management plan occurs in a context of significant efforts by other organizations to provide portals, catalogs, and gateways to environmental and atmospheric data resources: for example, NASA's Global Change Master Directory, the ARM Programs's Data Archive, the U.S. Global Change Research Program's Gateway to Global Change Data, NOAAServer's access to distributed data, The DoD's Master Environmental Library (MEL), the Earth Observing System Data Gateway, DOC's National Environmental Data Index, and the Federal Geographic Data Committee's Clearinghouse. Links to these and others are available from <http://www.unidata.ucar.edu/staff/russ/dmvg/portals.html>.

A fundamental issue is how NOAA should interface with these other efforts. Two extreme approaches would be to 1) develop the necessary infrastructure to participate as an institution with a unified collection of data assets and services, or 2) continue as separate and nearly autonomous research groups, each with their own tailored data standards and policies for sharing data with specific communities.

The first extreme would standardize access to NOAA's diverse data assets and services by creating a single NOAA format and virtual data portal, independent of discipline boundaries or user requirements, making it possible to monitor usage, standardize protocols, and enforce metadata standards and policies on an institution-wide basis. Such an effort would develop an

organized data collection from numerous existing collections of observational data, model outputs, field experiments, and derived, value-added, data. It would require not only organizing existing data assets, but also constraining new data assets and services to fit within a NOAA-wide integrated data management standard framework that emphasized interoperability.

At the other extreme, which resembles the status quo, each NOAA group responsible for data assets or services would make these available individually to other appropriate organizations, disciplines, and users independently of other efforts at NOAA. Rather than choosing institution-wide standards for the representation of metadata or interfaces for data access, use of standards most appropriate for each specific data resource or service would be encouraged. For example, archived climate data would be made available through servers and application programming interfaces most familiar to climate researchers, whereas real-time data from a field experiment would be provided in a form most useful to the investigators involved with the experiment. The fact that NOAA-managed groups made both data sets available would not influence the technical decisions that determine how those data sets were organized.

The first extreme represents a level of centralization and standardization that is practically not achievable and not desirable. Constructing or selecting suitable standards to encompass all the data at NOAA is a difficult task; by the time all existing NOAA data collections were molded to make them conform to selected standards, those standards might be obsolete and the cost in terms of personnel time would very large. The provenance of data is an important data attribute, but it should not be the primary attribute determining the organization or representation of the data, especially when the effort to do so stifles innovation.

The second extreme, complete data autonomy, is also undesirable for several reasons. It requires every group and project in the organization responsible for data to design and implement their own means for making that data available to others, including awareness of the best practices for metadata representation for discovery and use, knowledge of how to make the data useful to a larger set of current and future uses than the specific project that generates the data, and resources for providing all the data services that are needed for efficient access to the data. Data autonomy and lack of resources lead to lack of awareness of beneficial connections among the data collections, and no way to readily determine how to access data from other groups in the organization. The scientific scope in research projects is broadening, not becoming more focused. To remain in concert with the scientific needs we must further integrate our available data resources.

Instead of either of these extremes, loosely combining legacy systems while developing new ways to support data access to NOAA data assets would permit NOAA to work on the cutting edge of distributed data systems. To achieve this interoperability, significant effort will be required from programmers and scientists throughout the organization. The system developed must be a benefit to the contributors as measured by improved service to their respective existing and growing data communities.

## **NOMADS Development Strategy**

The NOMADS will foster interoperability by integrating existing working systems, relying on local decisions about systems that have evolved to successfully occupy a data niche, rather than imposing from the top-down standards that may be inappropriate.

## Metadata Conventions

In its broadest sense, metadata are simply ‘structured data about data’, which describe attributes of an information resource. Everyday examples of metadata include such things as the advertisement, directions for use and the nutrition information panel on the side of a food product, the description of products in a mail order clothing catalog or the table of contents of a book. General metadata examples of greater interest to researchers would be descriptions of telescope images or the header files describing gridded model output.

Metadata can be conceptually classed into two general types, discovery and use. Discovery metadata addresses the information necessary to identify a data collection and determine its availability and appropriateness for the intended application. Use metadata provides the technical information necessary to actually use the data in the collection. Of the two types, use metadata are more mature due to the creators and consumers of geodata converging in the last decade to a modest number of data storage formats containing reasonably well defined data descriptions. Discovery metadata has only recently become an issue as operational and science centers have begun to move from static, in-house, data archives to more dynamic, online, data services.

Efficient exploitation of massive data sets requires cataloging and documentation through the use of metadata, i.e., data describing the primary data objects themselves. In addition, verifiability of simulation-based research requires systematic collection and maintenance of metadata that document the design and execution of a simulation or collection of simulations. Locating science information within the massive data archives is currently difficult and requires considerable intimate knowledge of the organization and structure of the archive. To facilitate discovery, metadata must be standardized and organized into databases that support a variety of query types. Different classes of queries require different types of metadata to identify information such as what data are available, the nature of the data, how they were generated, and where they are located. Current metadata conventions used in the community (COARDS, CSM, GDT, SOHO-FITS, CEDAR, etc.) address primarily the description of the contents of individual files. These conventions need to be extended to encapsulate information about data collections and their derivation history. For example, environmental simulation systems are often composed as distributed applications; each component can represent a physical subsystem, such as the atmosphere, the ocean, or a level in a grid hierarchy. Each component may be responsible for its own output processing. Metadata must identify the relationships between the components to allow reconstruction of the overall simulation configuration. Similarly, data will often pass through many post-processing steps after the completion of the simulation. Each of these steps needs to be documented in the metadata. Identifying the appropriate common semantics and granularity of discovery metadata, upgrading legacy use metadata for online applications and ensuring that metadata are retained in a dynamic environment are all topics which will need to be addressed to successfully implement location independent data services across distributed data

centers.

## **Areas of Development**

In order to achieve the goals advocated by this plan, NOAA data managers must direct their efforts towards expanding data services in several interconnected areas. In many cases, progress in these areas will require NOAA to coordinate with several groups such as NCAR, NSF and projects like the "Community Data Portal", graphics/visualization efforts in NOAA, NASA, NCAR, DOE, and UNIDATA real-time data services and others. Over the long term, thrusts in new directions, such as the integration of NOAA data into the GIS world or the linking of NOAA data services to national/international data repositories will require outside collaborations.

A continued effort to expose NOAA data handling staff to data management advances and issues within different divisions and groups are important. NOAA can initiate smaller task forces to look at specific issues of data services in more detail, which will speed up the implementation process. A continually open channel of communication between these groups will help lessen the burden of re-inventing services at the local levels, encourage partnerships between groups with similar needs and help the NOAA staff understand the issues and needs of the broad group of NOAA data handlers.

While the underlying connecting structure of these areas relies on the adoption of interoperable, flexible, metadata representations by NOAA data producers, other developments that need to be addressed by the NOMADS, in no particular order, include: DOE, NASA, DOA, EPA, and all other programs within NOAA.

## **Data Cataloging**

Common data services will rely on dynamic data catalogs generated by automated metadata parsers. Since the actual amount of metadata for different datasets can vary, it will be necessary to allow for different levels of populating metadata. The syntax and semantics of queries on such metadata catalogs must support searches based on science questions in addition to the more conventional content-based queries. It will be necessary to develop a coordinated approach to populating, using, maintaining, and presenting consistent metadata catalogs across the NOAA, university, and related communities.

Query results returned to the user must be organized in an ordered, schematic and hierarchical format. These results must include both local data sources stored at NOAA, and remote sources stored at universities and collaborating institutions around the world. Data cataloging must support a tree-like structure model, in which the user can rapidly move from a more general selection, such as a complete field campaign or a selected geographic region, to a very specific request like the humidity data collected by a certain instrument at some location on some day, to possibly the elemental datum component. Also, the data cataloging system must be dynamic, indexing both non-volatile and real-time data sets, often made available from remote locations, and must be able to handle large scale, distributed data archives.

Locating the requested data is only the first step in deciding whether or not those data will be relevant to the proposed scientific objectives. Data cataloging must be enhanced by a comprehensive, in-depth description of the data specifications such as data quality, data type, data medium, which need to be extracted from the corresponding metadata fields. These in turn can point to additional resources describing the data such as journal papers or web pages, or tools enabling the analysis and visualization of those data, for authentication or quality assessment.

Work in this area is already proceeding at several NOMADS collaborator's including UCAR, LLNL, the OPeNDAP Inc., and others. Unfortunately, too few developers are aware of these ongoing efforts. Some examples are:

- PyClimate, a freely available climate model diagnostics tool with low frequency and statistical tolls for climate related analysis;
- NCAR's freely-available relational database system for metadata, some object-oriented wrappers, and dynamic generation of web pages to provide access to specified subsets of CEDAR data through a Web interface from CEDARWeb.
- NOAA and NCAR's work with the Earth System Grid (ESG) project to represent metadata in a Lightweight Directory Access Protocol (LDAP) server using a simple text format; and
- Unidata's development of a platform-independent Catalog Server that connects applications with data collections whose contents are described in XML called THREDDS.

These four projects are using completely different ways to represent metadata for access by applications: a relational database, an LDAP directory and XML. The issues involved and the lessons learned from these approaches ought to be of interest and use to other efforts to make data widely available.

**Data Cataloging Strategy:** The NOMADS Technical Team will organize presentations of these and other approaches to serving metadata and to compare and contrast their potential benefits to other data projects with the goal of increasing knowledge about metadata services and constructing a prototype metadata portal.

## **Data Discovery**

Data discovery represents the connecting link between data collection and archiving by a research group and use of the same data for scientific research by another group. In an age of massive datasets and distributed information environments, data discovery is an indispensable tool for comprehensive, collaborative and innovative scientific research.

In a general schematic model of a data discovery service provided by UCAR, a local or remote user would use a client application such as a web browser, some desktop program or a wireless device to submit a personalized data query to a NOAA data portal. The request would trigger

the action of a dedicated search engine connected to a global catalogue database, populated with the discovery metadata associated with all enlisted datasets. When the query is finalized, the returned results point to one or more well defined datasets. These can be either the full physical dataset stored locally or remotely in one of a wide variety of formats, or subsets of the original dataset, or "virtual" datasets obtained by some manipulation of the variables contained in the original dataset.

The data query interface presented to a remote user must be widely accessible, easy to use, and extremely flexible. Users must be able to connect to the world wide web and submit a data query in a wide variety of formats: using pre-defined keywords, free-form expressions, space and/or time specifications, and possibly SQL statements. Also, the query syntax must support the discovery of relevant derived variables obtained from the geo-physical variables actually contained in the datasets (for example, a query for wind speed  $\sqrt{u^2+v^2+w^2}$  when the dataset contains the u, v, w components of the wind). Advanced web software technology supporting session-state tracking and database connectivity, in conjunction with image displaying tools, may be employed to guide the user through a short series of steps to efficiently and rapidly locate the required data. As an additional requirement, the data query interface must rely on an underlying well defined interface that can be easily accessed by both human users and automatic computer programs. NOAA's Live Access Server (LAS) interface provides an implemented example of a data discovery system that already fulfills many of these requirements.

## **Data Access**

The data providers at NOAA have an important role in an overall data services strategy. Their mission is to provide high quality data and insure that these data are visible and available to a large research community. By taking advantage of existing mechanisms and technologies to make their data accessible, they are better serving the community. By developing useful metadata using widely-used conventions, they are providing valuable information to the researcher, which can be found using advanced tools. In return, a data service layer must provide value-added capabilities to the data providers by facilitating discovery and providing a common interface to similar data within the organization. In this way, the scientific community benefits and the data from the provider attain wider use.

In many cases, a data provider may experience significant advantages by providing their data using a client/server data access model instead of a file based model. A primary advantage of this model is that it facilitates access to datasets that is separate from its physical location. A client/server model provides a higher level of separation between the data layer and the application layer. An application does not need to know the details of file formats, low-level platform-dependent I/O system calls or the file names and locations. The application can access small subsets of large datasets without consuming large amounts of local disk space. Finally, the application will automatically be informed of data updates without having to know about the file-level changes of these updates. Of course, the client/server model is heavily reliant on the network and the data server's ability to appropriately present the data using well-understood conventions and useful mechanisms for subsetting and format conversions. In some cases, it may

be less efficient to use the client/server model when data are accessed many times or by many applications and for which it makes more sense to handle the data locally.

It can be extremely beneficial for the data provider to use well-developed technologies such as LAS, OPeNDAP, ADDE and IDD to more quickly and easily tap into methods of distributing their data over the Internet. In all cases, the network connectivity and bandwidth is crucial to this process. In the near future, technical advances in remote satellite uplinks can offer Internet and data distribution for data providers operating in remote locations; NOAA should invest in these technologies to facilitate remote data distribution. Finally, this continued and increasing reliance on networking will require that this institution remain committed to supplying adequate resources to the NOAA networking group.

### **Data Archiving and Preservation**

The future development and deployment of the NCDC archive is bracketed by both the need to continue providing traditional mass storage file storage and access services in the context of a full 7 by 24 operational production environment, as well as adapting the archive to new roles as an integrated component of larger data management efforts NOAA-wide.

The traditional role of reliably preserving NOAA's data and serving them to a wide client base is critical and should continue. The NCDC archive must scale up to meet an ever-growing demand for secure data storage and high-performance access without human intervention. This entails the constant evaluation and periodic deployment of the latest, highest performance, and most cost-effective hardware and software technologies available for peer-to-peer connectivity.

Finally, an ever-growing repository of data files requires better management tools, both for archive administrators and for individual users of the archive. NCDC must leverage Information Technology advances in areas such as database, data warehousing, analytic processing and data mining, statistics and visualization to record and analyze the creation and subsequent usage of archive data objects. Internally, this activity will support more effective operations, capacity planning and the management of what is expected to be explosive future growth. Externally, metadata and accounting tools will allow users to more conveniently and cost-effectively manage their own MSS data assets. These tools are all expected to be Web-enabled in the future.

## **Data Coordination and Training**

User education is an important piece of any strategy for improved data services and interoperability. The NOMADS data management process must include the education of all data providers and data users about how to go about presenting their data in a form that is compatible. We see three main areas of education and training needed:

### **Data Providers**

The scope of data providers at NOAA is very large. Output from computer models, satellite imagery, real-time radar and weather station feeds and aircraft are just a few examples of the variety of data available at NOAA. Over the years, the role of data providers at NOAA has been primarily of working with a fairly specific suite of datasets in a specific discipline and developing analysis software for these data. In addition, data providers spend increasingly more time managing these data as size and bandwidth increases. In an interoperable world, the role of the data provider must also now include attention to data services.

### **Data Service Developers**

A relatively new area in software development for most in NOAA is Data Service Layer software. Education, seminars and workshops discussing the topics of software frameworks and other current software engineering concepts will benefit our software engineering efforts here. In order to develop this middle tier, software developers will need to be educated in the best methods to provide this service layer to the large variety of data formats at NOAA. This will likely include database technology and distributed access methods. For retrospective data, knowledge of NOAA's vast archival capability will be important.

In addition, API level education and a more intimate knowledge of the types of functions and software methods that are needed by higher-level applications developers is paramount. Many of the existing training courses have been limited to user-level training in the latest Microsoft applications software, for example. We will need to emphasize software engineering, databases, metadata and design in all aspects of training in order to increase coordination and communication among software engineering staff in order for this development to succeed.

### **Display and Analysis Tool Developers**

There are a myriad of software tools developed at NOAA, many of which overlap in functionality. In order to make the most efficient use of software engineering talents, a common data access layer can allow NOAA data users to view and intercompare our data using a well understood set of tools. In many cases, these analysis packages will eliminate the need for developers to "re-invent" software at the local level since it is already available from other sources. This can allow local software developers to place more emphasis on, for example, automated data quality algorithms and other mechanisms to improve data quality. Education in this area takes the form of keeping the software developers aware of what is already available as well as training in the data access methods of the data service layer for successful software

development.

## **Data Coordination and Training Strategy**

The NOMADS team will act as the forum through which data managers, data service providers and data tool developers can coordinate their efforts to develop a comprehensive common data access infrastructure.

## **Implementation**

Implementation of this plan will begin along two lines. The first is formation of the NOMADS Steering Group, a Science Team and a Technical Team. The Technical team will research and identify a minimal set of metadata elements; establish a prototype data portal based on existing cross-cutting data management services; and determine compatibility and support to existing efforts such as OPeNDAP. The Science Team will conduct requirements analysis and provide recommendations for long-term archive of selected data to the NOAA Archive Board; while the Steering Group will provide program oversight, long term vision and develop funding strategies for NOAA Management.

### **The NOMADS Technical Team: Metadata**

One of the first tasks of the NOMADS Technical Team will be to examine the existing metadata standards and assess their applicability for NOAA data needs. Those standards that show promise will be candidates for prototype XML-based translators into the NOMADS framework. While the Technical Team will not force any data providers to choose a specific metadata format, those data managers who chose a metadata standard for which a NOMADS convention already exists will avoid having to develop their own conventions.

Existing standards for interoperability of metadata include Dublin Core, FGDC, DIF, and GILS. Rather than formally adopting one of these standards, the Technical Team will encourage NOAA data managers to systematically use some metadata representation. Formal metadata use will allow for automatic translation to XML, since XML as a syntax for representing arbitrary metadata subsumes the other standards.

The NCDC Mass Store Archive, for example, has maintained metadata in FGDC standards in text form for several decades. Currently, the text forms are automatically manipulated to create a HTML documentation system. A similar strategy could be employed to represent these metadata in an XML standard so that applications and search engines could equally discover these data along with other NOAA data sources represented by the same standard.

### **Prototype Data Portal**

In the near future, it will become difficult, if not impossible, for data providers themselves to develop applications and infrastructure to allow their data to be used by the wider scientific community that the NOAA is directing us to serve. Emphasizing the use and development of

existing data services and archiving infrastructure (e.g. OPeNDAP, LDM/IDD, LAS, MSS, GrADS-Data Server, etc.) will become a critical component of success. Our organization can benefit by selecting, enhancing and encouraging the use of a data services layer that can encompass as many of these data sources as possible.

In the long term, since the data management goals and needs are common to the whole geophysical community at large, it will be extremely important that the NOMADS plan be orchestrated on two levels: first, by coordinating all existing and forthcoming efforts inside NOAA (CLASS, NVODS, etc.); and second, by collaborating with all other scientific institutions funded by NOAA, DoE and NASA towards a common environment of distributed data sharing and visualization.

### **OPeNDAP Overview**

In the early 1990's, a workshop consisting of approximately 40 oceanographers, computer scientists and oceanographic data archivists from Government and academia, research and operations met to design a data system that would facilitate the exchange of data over the Internet between research scientists, federal archives, and private organizations. This group in effect was addressing many of the issues related to Task A1 of NOPP BAA 99-027. Two fundamental design criteria emerged from the workshop: (1) servers must be easy to install, and (2) the system must interface to existing application software. A system that satisfies this basic philosophy will allow individual researchers as well as national archives to be data providers, and it will allow researchers, operational modelers, interested hobbyists, ...everybody to use familiar and appropriate software.

OPeNDAP is the system that was designed to satisfy these criteria. Implementation of the basic system was completed two years ago and, with funding from NASA, NOAA and NSF, the system has been significantly enhanced over the past two years.

The OPeNDAP approach is to use the standardized interfaces defined by multiple file API's (e.g. NETwork Common Data Format (NetCDF), HDF, ...) as the point at which to insert the distributed data infrastructure. In this approach existing applications - both commercial applications and those built within the science community - are "relinked" with new libraries that masquerade as the original file I/O library. The applications are unaware that they have been extended to perform network access. The data from remote files are made available through servers that invert the process - using the standard file or data base APIs to read the files and then provide the data over the Internet in a format-neutral representation. The virtues of this approach are adaptability, leveraging, and invisibility:

1. The investment that each scientific project has made in its software tools is protected. Users continue to use the software tools with which they are already familiar - now extended to perform remote data access.
2. The approach leverages hundreds of already-existing low-level file manipulation utilities. For example, utilities such as the NetCDF Operators ("NCO") which subset, reorder, and append data from files immediately become network tools for performing the same operations on

widely-distributed data sets.

3. Format-independence is achievable through this approach. Applications communicate with files through standardized interfaces (API's), without knowledge of what occurs behind those interfaces. Format translation may occur without the application being aware of it.

### **The OPeNDAP Organizational Structure of the Partnership; Project Management**

Overall project direction will be provided by an Executive Committee (ExCom) selected from the Primary Partners. These individuals are: Cornillon (Chair; URI), Flierl (MIT), Gallagher (URI), Hankin (NOAA), Mercer (State of Maine), Nowlin (Texas A&M) and Chinman (ex officio; UCAR).

The partnership is organized into two levels: a regional level and a national/international level. Partners participating at the national level are project Primary Partners (CO-Investigators or Co-Is). Partners participating at the regional level are Secondary Partners. Five regions defined for the continental US make up the regional level: the northeast (Mercer), the southeast (Davidson), the Gulf Coast (Nowlin), the west coast (Abbott) and the Great Lakes (Andren). Each region will be lead by a regional coordinator who will also participate at the national level hence is a Primary Partner. Names in parenthesis above are the regional coordinators. To insure a broad range of inputs from the five regional consortia, regional coordinators with very different backgrounds were purposefully selected.

The national/international level will consist of the regional coordinators, representatives from federal agencies with oceanographic data holdings, private industry, individuals representing national organizations of marine institutions, individuals with ties to international organizations and individuals with data system/network experience.

In addition to the proposal Co-Is, representatives from the other NOPP funded activities with a VODHub connection will be invited to participate at the national level. Finally, Ron Baird of the National Office of Sea Grant, Steve Hale of EPA, John Lever of NAVOCEANO, and Hank Frey of NODC have indicated a willingness to participate at the national level, the last two if this effort is the sole funded respondent to Task A1 (NAVOCEANO and NODC are submitting a competing proposal). Admittedly, the connection with NODC is not well defined given Frey's retirement.

The Principal Investigator of the project is P. Cornillon of the University of Rhode Island. Gallagher (URI) will be the technical lead for the project. Chinman (UCAR) will serve as project manager. Dan Holloway will be responsible for coordinating server installation. Holloway (URI) will also have overall responsibility for user interface issues while Hankin (Pacific Marine Environmental Laboratory (PMEL)) will lead the web interface work. Hankin, in coordination with Olsen at GCMD, will lead the data discovery portion of the project. Habermann (National Geophysical Data Center (NGDC)) will coordinate the GIS work. Each of the regional coordinators will be responsible for regional population and Cornillon will coordinate the national population effort.

## **Interoperability and the OPENDAP Three Tiered Data System**

Earth science data systems are generally viewed as consisting of three primary levels:

- the Directory Level provides a list of datasets along with the parameters available and the approximate temporal and spatial coverage for each data set. An example of an entry in such a system is the hydrographic data set at the National Oceanographic Data Center (NODC);
- the Inventory Level provides a detailed listing of the data granules within a data set. For the NODC hydrographic data set, this might consist of a listing of each cast along with the location (latitude and longitude) and time of the cast;
- the Data Level consists of the actual data objects.

Directories have generally been maintained separately from inventories and from the actual data, while inventories, when they exist are often found co-located with the data although inventory access functions and data access functions have been kept separate.

This hierarchical view has resulted from practical considerations related to networking and storage capacity in the past and, now that these constraints are relaxed, other structures are practical. For example, directory and inventory information could be combined and maintained at the same sites as the data in a totally distributed system, one in which the directory and inventory functions, and possibly the data acquisition function are combined. There would be no distinction between the levels in such a system. For the purposes of this discussion, we will maintain the three layer historical view: directory - inventory - data. We believe that this will help to better understand the issues involved, but we want to emphasize that we are not "married" to such a hierarchical approach (more on this later).

The degree to which system-wide interoperability is achieved depends on the metadata associated with the data. Metadata is information about the data. Generally, when metadata are discussed, one is referring to information about the contents of the data; e.g., the variable T refers to sea surface temperature and its units are degrees Centigrade or the data set covers the period 8 January 1982 through 29 May 1990. We prefer to take a broader view of metadata, dividing it

into two basic groups syntactic and semantic metadata. Syntactic metadata is information about the data types and structures at the computer level, the syntax of the data; e.g., variable T represents a floating point array measuring 20 by 40 elements. This is information that is required as part of the transport protocol for the data in a network based data system. Semantic metadata is what one normally thinks of as metadata, information about the contents of the data set.

### **Syntactic Metadata: The Data/Metadata Transport Protocol**

Without a rigorous syntactic description of the data that are being moved from one place to another it is virtually impossible to make use of them, the data stream is simply a large collection of bytes. We therefore begin the metadata discussion with a characterization of this metadata type.

The data transport protocol is characterized by the data model, the organizational description of the data as they are moved between client and server<sup>6</sup>. The data model generally consists of data types, e.g., byte, integer, string, etc. and groupings of these data types, e.g., arrays, lists, etc. html is effectively a data model albeit a very simple one consisting of string data, metadata in the form of mark-up tags, and meta-data indicating inclusion of external (opaque) content, e.g., GIF images. The Hierarchical Data Format (HDF) is a much more sophisticated data model designed primarily for array data although it has evolved to include sequences as well as complicated data structures. The Distributed Oceanographic Data System (DODS/OPENDAP) data model achieves the generality to encompass a range of such underlying models through its extensibility - complex structures may be assembled from more basic structures. It has been designed explicitly for oceanographic data and is used only as a transport protocol, not as a storage format as is HDF. The OPENDAP data model consists of data types: Byte, Integer, Short Integer, Float, String and URL and groupings of these data types: Array, Structure, Lists, Sequences and Grids. The OPENDAP User's Guide<sup>7</sup> contains a thorough explanation of these data-types and type groupings.

In addition to data types, the data model may also be considered to include operations that may be performed on the data such as subsetting, projection, etc. For HDF these functions are part of the Application Program Interface (API). In OPENDAP they are part of the operations permitted by OPENDAP servers.

Regardless of the data system level considered there are two ways to achieve interoperability with regard to the transfer of data in a distributed data system: (1) require that all of the data are stored in the same format (data model) in which case the transport data model is the same as the data model used to store the data or (2) translate from the format in which the data are stored to the format expected by the application requesting the data. The latter can be done in one step, every format is translated to every other format (i.e.,  $n(n-1)$  translation functions) or in two steps, the data are transformed to an intermediate format and then to the final format ( $2n$  translation functions). Regardless of the approach taken, the data model must be rich enough to accommodate all of the data types encountered in the system. OPENDAP makes use of the second approach, translates to an intermediate data model and then to the applications data model.

In general, the complexity of the data model increases as one moves from the directory level to the data level. At the directory level the data model need not be more complicated than that used for html while at the data level, html will clearly be inadequate. This means that if the data model adopted is sufficiently rich to accommodate the actual data, it should also be able to accommodate information at the inventory and directory levels.

### **Semantic Metadata**

To facilitate the use of data one requires semantic metadata, information about the data. In the discussion of semantic metadata it is useful to distinguish between metadata required to use a

data set, i.e., the semantic metadata transmitted at the data level, and that required to find a data set containing data of potential interest, i.e., the metadata used at the directory level. We refer to the first as use metadata and the latter as search metadata. This distinction is quite important because most metadata discussions center around search metadata requirements and not use metadata requirements; e.g., the Directory Interchange Format (DIF) of the Global Change Master Directory (GCMD). There is overlap between use metadata and search metadata, but one is not a subset of the other; e.g., to search for a data set, missing value flags are not required while to use the data such information is crucial. Similarly the range of the variables in a data set is not required to use the data but forms the basis for many data set searches.

Use metadata may be further subdivided into translational and descriptive use metadata. The former refers to operations that are performed on the data values, be they the names of the variables or the digital numbers associated with them, that are required for the user to understand their meaning. For example, the variable  $T \rightarrow \text{sea surface temperature}$  or  $d \times 0.125 \rightarrow ^\circ\text{C}$ , where  $d$  is the number stored in the data set. Descriptive use metadata is information about the data such as how the instrument was calibrated or what sensor was used.

Search metadata may also be further subdivided, in this case into parameter, range and descriptive search metadata. Parameter search metadata contains the list of parameters or variables in the data set. This could be further subdivided into dependent and independent variables. Range search metadata contains the ranges of variables within the data set. In most existing directory systems, only the ranges for time and space are included. Descriptive search metadata contains other information associated with the data set such as a generic description of the sensor used. There may be overlap between this descriptive information and that contained in descriptive use metadata, although this need not be the case. For example, a description of the sensor may be relevant to both groups, but there is no reason to include detailed information about sensor calibration as descriptive search metadata.

Figure 1. shows the different metadata types schematically. Although three levels are shown in this figure, the inventory and data levels have been treated together. This is the approach that has been taken in OPENDAP; i.e., inventory information is treated in the same way as data. Interoperability at the data level with inventory access requires those squares with green and magenta backgrounds. The green square requires a rigid metadata description, while the magenta squares need not be as rigid. Descriptive use metadata is not required for interoperability at the data level. The gold squares denote metadata that is required at the directory level for system wide interoperability, to search for a data set the system requires access to knowledge of the variables in each data set as well as the ranges of at least some (space-time) of these variables.

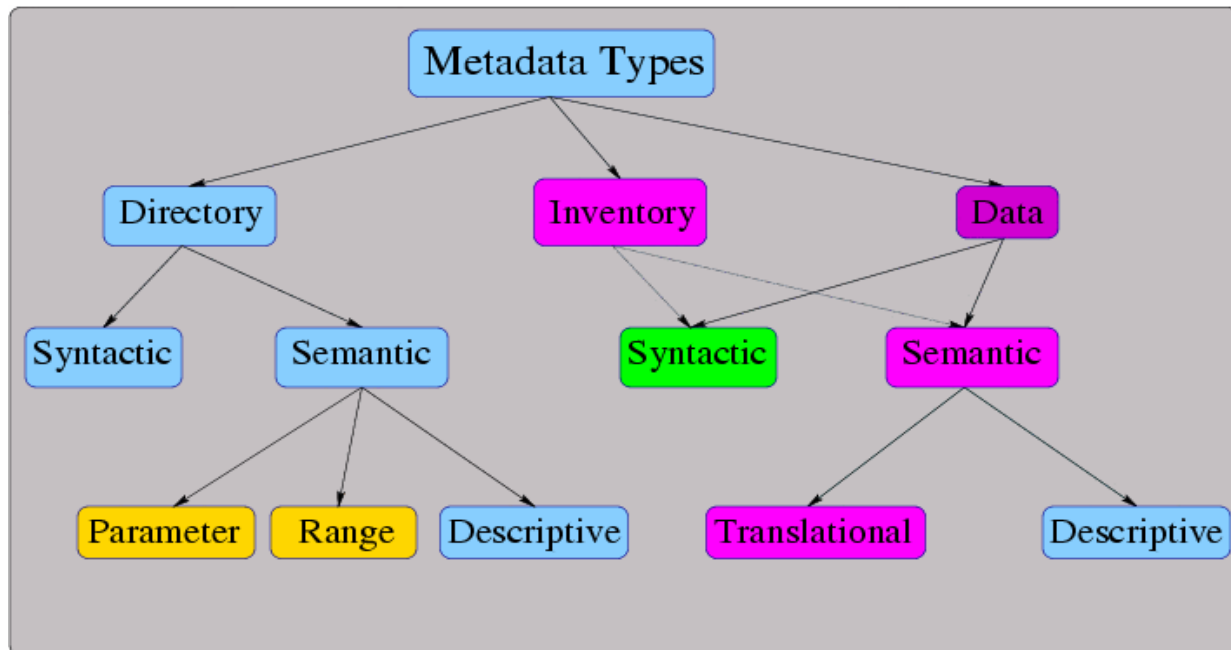


Figure 3.

A categorization of metadata in a three level data systems.

Again descriptive metadata is not required at the search level although it may be useful. Demands for semantic metadata are among the more contentious issues that one faces in the design of a distributed data system. Users generally want more semantic metadata than providers are interested in generating. For example, the GCMD found that when too much search metadata was requested, many data set generators would simply not provide data set descriptions to the GCMD. The GCMD developed the "skinny DIF" to address this problem. The "skinny DIF" is a data set description that contains the minimum number of fields thought to be necessary to satisfy a rudimentary search.

### The Data Access Protocol

The DAP is a simple request/response protocol which provides access to data. The DAP is used to transfer three types of objects from remote sites (servers) to a client: 1) An object which describes the variables contained in a dataset by listing their name, data type and relationship with other variables in the dataset, 2) An object that lists attributes for each of those variables and, 3) A data object which contains both variable names, types and relationships along with numeric values for each variable. A request for a data object can be accompanied by a Constraint Expression (CE) which lists which of the variables, and of those variables which values or ranges of values, to return. Thus the DAP provides both a simple way to ask a dataset about its contents and to selectively read from that dataset.

The DAP has been implemented in two different languages, C++ and Java. The C++ version is available on a variety of UNIX platforms. In addition, this software has been ported to the Microsoft Windows operating systems (Windows 95, 98 and NT) and is now in beta test. The Java implementation of the DAP will be completed in January 2000.

There are some significant enhancements to the DAP on which we and others are currently working. These include server-based data type translation, geographical selection functions, response caching, greater control over the transmission process, better error handling and more efficient memory management. None of these modifications alters the DAP in a significant way but they will all result in either simplified use of the system or in improved system performance. Most of the enhancements were initiated by the user community. They will be completed with current funding.

### **OPENDAP Client and Server Software**

The OPENDAP project software most visible to users are its clients and servers. OPENDAP clients fall into three categories: client-libraries, command line clients and Graphical User Interface (GUI) clients. The client-libraries are software libraries that can be used by programmers and sophisticated users to build clients for OPENDAP with existing software. These client-libraries implement already existing APIs but do so in a way that enables them to read information from OPENDAP servers. Four client-libraries are currently available: NetCDF 3.x, JGOFS, Java (native) and C++. An example of the use of the NetCDF client-library has been its linking with Ferret9 and with GrADS10 to access OPENDAP servers. Client-libraries are a powerful software tool because they make it simple to leverage existing work in the important process of building a useful body of client software.

Currently available OPENDAP command-line clients work with the UNIX (and soon Windows) operating systems and with the Matlab, EXCEL and IDL commercial analysis packages. These clients require users to compose OPENDAP URLs, which many find a complex task, but provide a useful basis for testing servers, and more importantly, writing scripts which read from OPENDAP servers. The command-line clients for both Matlab and IDL are similar in function to the client-libraries because they provide a programmatic interface to OPENDAP servers. The Matlab and IDL command-line tools can (and have) been used to build complex graphical interfaces and scripts which automate access to data served by OPENDAP under program control.

OPENDAP Graphical User Interfaces exist for Matlab, IDL and Ferret. These interfaces greatly simplify the specification of subsets of the desired data; the user selects data ranges in a point-and-click interface and the Graphical User Interface (GUI) builds the Uniform Resource Locator (URL), issues the request for data and scales the data on return if desired.

OPENDAP currently provides six different servers<sup>11</sup> which make it possible for clients to access data stored in many different ways. The FreeForm and Joint Global Ocean Flux Experiment (JGOFS) servers were built because they are easily customizable to a wide variety of simple file formats that are often used by the scientist and, more recently, by the national archives for the data that they make accessible via their web sites.

One of the more serious problems faced in developing a data access system relates to how the system handles multi-file data sets. As presently conceived OPENDAP servers such as those for HDF or NetCDF generally treat each file in a data set separately; i.e., each file is accessed via a separate URL. (This is not true of the JGOFS servers.) This is to some extent imposed on

OPENDAP by the way the HDF and the NetCDF API's treat multi-file systems. The first step in addressing this problem has been to develop a file server. This is a OPENDAP server that provides ordered lists of URLs based on the characteristics or coordinates that differentiate the individual files in the dataset; e.g., time. Accessing data from such a data set generally involves a two step process: first a request is made to the file server and then, using the results of this request, other requests may be made for the actual data values. In the next year the OPENDAP core group will extend this capability so that accessing the catalog and accessing data in any of the files it lists can be made in a single step, effectively aggregating the files into a single dataset from the perspective of a client program.

### **The World Wide Web and OPENDAP Software**

Thus far we have described OPENDAP in terms of its abstract architecture: connecting existing clients to existing data formats by means of the DAP. The DAP is, in fact, a distributed object model. Deployment of the DAP over real networks requires the use of a network transport layer. The DAP is compatible with most modern network transport and distributed object protocols: http, CORBA/IIOP, Java/RMI, etc.. In fact, it is reasonable to think of future implementations of this approach which utilize multiple protocols, much as today's acoustic modems select among available protocols.

The current implementation of OPENDAP is layered on http, which offers the following practical advantages:

- 1.http servers are generally already installed at potential data provider sites,
- 2.expertise installing and managing http servers is available at most potential data provider sites,
- 3.firewalls are typically configured to let http pass through, and
- 4.there are typically no licensing barriers associated with http implementations.

Furthermore, because OPENDAP is currently layered on http any web browser can be considered a OPENDAP client. We have capitalized on this in several ways. OPENDAP servers support a variety of http/WWW based accesses.

1. All of the basic objects returned by servers in response to a request are pseudo-MIME documents. The objects describing variables and their attributes are text documents and can be displayed by any web browser. The data object is a binary document and browsers will offer to save it to a file.
2. It is possible to request data in ASCII form from a OPENDAP server. This information is returned as a plain text document which can be viewed in a browser, saved to a file or read by a spreadsheet. All the constraint expression features can be used in an ASCII request.
3. OPENDAP servers can also supply a simple html form-based query interface. This interface allows the user to view the variables and variable attributes for any OPENDAP dataset, to define a subset of that dataset by filling in entries on the form and to request the subset. The form will build the URL.

